

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



وزارت علوم، تحقیقات و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)

طرح پژوهشی

تحلیل رفتار جستجوی اطلاعات پژوهشگران

در موتور جستجوی سامانه‌ی گنج

مجری

سمیه فتاحی

۱۳۹۶

حقوق: پژوهشگاه علوم و فناوری اطلاعات ایران

طرح پژوهشی تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج پیرو قرارداد شماره ۳۳/۱۲۰۰۷ مورخ ۱۳۹۵/۹/۳۰ میان پژوهشگاه علوم و فناوری اطلاعات ایران (کارفرما) و خانم دکتر سمیه فتاحی اجرا شده است. گزارش حاضر یک جلد از مستندات این طرح است.

این گزارش و تمامی حقوق مادی آن بر اساس قانون حمایت حقوق مولفان و مصنفان و هنرمندان، مصوب ۱۳۴۸ و اصلاحیه‌های بعدی آن و همچنین آیین‌نامه‌های اجرایی این قانون متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران و هرگونه استفاده از تمامی یا پاره‌ای از آن، شامل: نقل قول، تکثیر، انتشار، کاربرد نتایج، تکمیل و مانند آنها به صورت چاپی، الکترونیکی یا وسایل دیگر فقط با اجازه کتبی پژوهشگاه امکان‌پذیر است. نقل قول در حد هزار واژه در انتشارات علمی مانند کتاب و مقاله با درج اطلاعات کامل کتابشناختی، نیازی به مجوز پژوهشگاه ندارد.

صحت مندرجات گزارش بر عهده مجری طرح پژوهشی است.

صفحه تاییدیه

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج

مدیر طرح پژوهشی: دکتر سمیه فتاحی، پژوهشگاه علوم و فناوری اطلاعات ایران.

نشانی: تهران، خیابان انقلاب، چهارراه فلسطین، شماره ۱۰۹۰، طبقه ۴، اتاق ۴۰۸

صندوق پستی: ۱۳۱۵۷۷۳۳۱۴ تلفن: ۰۲۱-۶۶۴۹۴۹۸۰ (داخلی ۴۰۸)

وبگاه: <https://irandoc.ac.ir/u/s-fatahi>

رایانامه: fatahi@irandoc.ac.ir

چکیده

یکی از رسالت‌های مهم هر سامانه‌ی اطلاعاتی ارائه بهترین خدمات متناسب با نیازهای کاربران می‌باشد. این امر باعث افزایش رضایت کاربران و علاقه‌مندی آنان به استفاده از سامانه‌های اطلاعاتی می‌گردد. امروزه موتورهای جستجو یکی از مهمترین ابزارهای بازیابی اطلاعات و بستری برای عرضه‌ی انواع اطلاعات برای بسیاری از کاربران هستند. دانشجویان، اساتید دانشگاه و پژوهشگران به طور روزانه به جستجو در موتورهای جستجو پرداخته و نیازهای اطلاعاتی خود را برآورده می‌سازند. پایگاه اطلاعاتی گنج (گنجینه علمی ایرانیان) پژوهشگاه علوم و فناوری اطلاعات ایران، با برخورداری از ۸۷۰ هزار رکورد علمی، امکان جستجو برای پژوهشگران را در پایان‌نامه‌ها، نشریات علمی داخلی، مقالات، همایش‌ها، طرح‌های پژوهشی و گزارش‌های دولتی فراهم می‌کند. واضح است که کاربران تحت تأثیر عوامل مختلف از شیوه‌ها و روش‌های گوناگون برای کسب اطلاعات استفاده می‌نمایند و رفتارهای جستجوی اطلاعات متفاوتی را از خود بروز می‌دهند. هدف اصلی پژوهش حاضر، تحلیل رفتار جستجوی اطلاعات کاربران در موتور جستجوی سامانه‌ی گنج می‌باشد. تحلیل رفتار جستجوی کاربران می‌تواند به درک بهتر نیازهای آنان کمک کند. نتایج تحلیل بیانگر آن است که میزان مراجعه حضوری در مقایسه با غیر حضوری بسیار اندک است و ساعات مراجعه حضوری در بین پژوهشگران حوزه علوم انسانی و فنی مهندسی متفاوت است. همچنین نتایج این تحلیل می‌تواند جهت ارزیابی و رفع نواقص احتمالی موجود در زمینه بازیابی اطلاعات توسط موتور جستجو کمک نماید.

کلیدواژه‌ها: رفتار جستجوی اطلاعات، تحلیل رفتار، موتور جستجو، سامانه‌ی اطلاعاتی، سامانه گنج

فهرست مطالب

۱	فصل اول - کلیات
۲	۱-۱- مقدمه
۲	۱-۲- اطلاعات
۳	۱-۲-۱- جستجوی اطلاعات و رفتار جستجوی اطلاعات
۴	۱-۲-۲- مدل‌های جستجوی اطلاعات
۴	۱-۲-۲-۱- مدل جستجوی اطلاعات ویلسون و کیر کلاس
۴	۱-۲-۲-۲- مدل فرآیند جستجوی اطلاعات بلکین
۵	۱-۲-۲-۳- مدل شش مهارتی ایزنبرگ و برکویتز
۶	۱-۲-۲-۴- مدل جستجوی اطلاعات الیس
۶	۱-۳- جستجوی اطلاعات در اینترنت
۸	۱-۴- تحلیل لاگ
۹	۱-۴-۱- مزایا و معایب جمع‌آوری و تحلیل داده‌های لاگ جستار
۹	۱-۴-۲- محدودیت‌های تحلیل لاگ‌ها
۱۰	۱-۴-۳- تقسیم‌بندی جستارها
۱۳	۱-۴-۴- سوالات متداول در پژوهش‌های بررسی لاگ‌ها
۱۴	۱-۵- نتیجه‌گیری
۱۵	فصل دوم- کارهای پیشین
۱۶	۲-۱- مقدمه
۱۶	۲-۲- تحلیل لاگ موتورهای جستجوی پایگاه‌های غیر علمی
۳۳	۲-۳- تحلیل لاگ موتورهای جستجوی پایگاه‌های علمی
۳۸	۲-۴- نتیجه‌گیری
۳۹	فصل سوم- روش انجام پژوهش
۴۰	۳-۱- مقدمه
۴۰	۳-۲- روش پژوهش
۴۶	۳-۳- نتیجه‌گیری
۴۷	فصل چهارم- نتایج و تحلیل آن‌ها
۴۸	۴-۱- مقدمه
۵۰	۴-۲- تحلیل داده‌ها براساس زمان و تاریخ
۵۵	۴-۳- تحلیل داده‌ها براساس کلیدواژه‌ها
۶۱	۴-۴- تحلیل داده‌ها براساس کلیدواژه‌ها و زمینه‌های علمی
۶۳	۴-۵- تحلیل داده‌ها براساس خوشه‌بندی کاربران

۷۲	فصل پنجم - جمع‌بندی و کارهای آتی
۷۳	۵-۱- نتیجه‌گیری
۷۳	۵-۲- کارهای آتی
۷۶	فهرست منابع

فهرست شکل‌ها

- شکل ۱- نمایی از سامانه گنج ۴۰
- شکل ۲- نمایی از لاگ موتور جستجوی گنج ۴۱
- شکل ۳- جستار جهت شمارش تعداد کل رکوردها ۴۲
- شکل ۴- جستار جهت شمارش تعداد رکوردها در بازه‌های جداگانه هر ماه ۴۲
- شکل ۵- جستار جهت شمارش فرکانس عبارات جستجو شده در سامانه گنج ۴۳
- شکل ۶- جستار جهت شمارش فرکانس عبارات جستجو شده در ساعت ۱۲ تا ۲ بعدازظهر در سامانه گنج ۴۳
- شکل ۷- جستار جهت استخراج تعداد جستجوها براساس روزهای هفته ۴۴
- شکل ۸- جستار جهت استخراج تعداد جستجوها براساس ساعات ۴۴
- شکل ۹- جستار جهت استخراج کلیدواژه‌های پرتکرار ۴۵
- شکل ۱۰- نسبت جستجوهای داخل، خارج از کشور و داخل پژوهشگاه ۴۸
- شکل ۱۱- میزان کل جستجوها به تفکیک روزهای هفته ۵۱
- شکل ۱۲- میزان جستجوها از داخل پژوهشگاه به تفکیک روزهای هفته ۵۱
- شکل ۱۳- حجم جستجوی انجام شده بر روی سامانه‌ی گنج به طور کلی براساس روزهای هفته ۵۲
- شکل ۱۴- حجم جستجوی انجام شده بر روی سامانه‌ی گنج از داخل پژوهشگاه براساس روزهای هفته ۵۲
- شکل ۱۵- حجم جستجوی انجام شده بر روی سامانه‌ی گنج از داخل کشور براساس روزهای هفته ۵۳
- شکل ۱۶- حجم کل جستجوها براساس ساعات شبانه روز به تفکیک ماه ۵۴
- شکل ۱۷- حجم جستجوهای داخل پژوهشگاه براساس ساعات شبانه روز ۵۴
- شکل ۱۸- مقایسه جستجوهای داخل پژوهشگاه و داخل کشور براساس ساعات شبانه‌روز ۵۵
- شکل ۱۹- بیشترین واژه‌های مورد جستجو در سامانه‌ی گنج در مهرماه ۵۶
- شکل ۲۰- بیشترین واژه‌های مورد جستجو در سامانه‌ی گنج در آبان ماه ۵۶
- شکل ۲۱- بیشترین واژه‌های مورد جستجو در سامانه‌ی گنج در آذرماه ۵۷
- شکل ۲۲- سهم جستجوی لغات در بین ۱۰۰ واژه پرتکرار در جستجوها ۵۷
- شکل ۲۳- آمار مربوط به جستجو برحسب ساعات مختلف جستجو براساس زمینه علمی ۶۲
- شکل ۲۴- آمار مربوط به جستجوی کلمات برحسب ساعات مختلف جستجو ۶۲
- شکل ۲۵- نمونه‌ای از لاگ‌های جستجو با id یکسان کاربران ۶۴
- شکل ۲۶- نمونه‌ای از لاگ‌های جستجو با id یکسان نشست ۶۴
- شکل ۲۷- نتایج خوشه‌بندی برای دو خوشه ۶۶
- شکل ۲۸- نتایج خوشه‌بندی برای سه خوشه ۶۶
- شکل ۲۹- نتایج خوشه‌بندی برای چهار خوشه ۶۶
- شکل ۳۰- نتایج خوشه‌بندی برای دو خوشه با حذف داده‌های کم تکرار ۶۷
- شکل ۳۱- توزیع کلیدواژه‌های جستجو شده برحسب طول ۶۹
- شکل ۳۲- مجموعه داده برچسب گذاری شده توسط خبره براساس ویژگی‌های مستخرج از لاگ ۷۰

فهرست جدول‌ها

۴۹	جدول ۱- میزان جستجوها توسط ip های داخل پژوهشگاه
۴۹	جدول ۲- میزان جستجو توسط ip های داخل کشور
۵۰	جدول ۳- آمار تعداد جستجوها به تفکیک ماه
۵۸	جدول ۴- لیست کلیدواژه‌های پرتکرار در رشته‌های مختلف در کل جستجوها به همراه تعداد دفعات جستجو
۵۹	جدول ۵- لیست کلیدواژه‌های پرتکرار در رشته‌های مختلف در جستجوهای داخل پژوهشگاه
۶۰	جدول ۶- لیست کلیدواژه‌های پرتکرار در رشته‌های مختلف در جستجوهای داخل کشور
۶۱	جدول ۷- آمار مربوط به جستجوی کلمات بر حسب ساعات مختلف جستجو
۶۵	جدول ۸- آمار تعداد جستارهای برچسب‌گذاری شده توسط خبره
۶۷	جدول ۹- آمار تعداد جستارهای برچسب‌گذاری شده توسط خبره پس از حذف برچسب‌های موضوعی کم تکرار
۶۸	جدول ۱۰- میانگین طول کلیدواژه جستجو به تفکیک زمینه علمی
۷۰	جدول ۱۱- نتایج خوشه‌بندی بر روی مجموعه داده‌ها شامل طول عبارات جستجو برای دو خوشه
۷۱	جدول ۱۲- نتایج خوشه‌بندی بر روی مجموعه داده‌ها شامل طول عبارات جستجو برای سه خوشه
۷۱	جدول ۱۳- نتایج خوشه‌بندی بر روی مجموعه داده‌ها شامل طول عبارات جستجو برای چهار خوشه
۷۴	جدول ۱۴- لیست کامل فیلدهای پیشنهادی جهت ذخیره‌سازی لاگ‌ها در سامانه گنج جدید

فصل اول

کلیات تحقیق

۱-۱- مقدمه

در این بخش مفاهیم مقدماتی در خصوص اطلاعات و تعاریف مربوط به آن، روش‌های جستجوی آن و رفتارهای مربوط به جستجوی اطلاعات، به ویژه در محیط اینترنت ارائه می‌شود. همچنین مفاهیم تحلیل لاگ به عنوان یکی از غنی‌ترین منابع ذخیره رفتار کاربران در زمان جستجوی اطلاعات بررسی خواهد شد. مزایا، معایب و چالش‌های پیش‌رو بحث خواهد شد.

۱-۲- اطلاعات^۱

امروزه، اطلاعات نقش مهمی را در نوع و کیفیت زندگی بشر بازی می‌کند و به عنوان منبع اساسی و مهم در فعالیتهای روزمره مطرح است و تمام تصمیم‌گیری‌ها تحت تاثیر اطلاعات است [۴۱]. اصطلاح اطلاعات، از دو کلمه‌ی لاتین "Forma" و "Formatio" گرفته شده است [۲۹]. هر دوتای این کلمات کم و بیش معنی یکسانی را از شکل دادن به چیزی و فرم دادن یک الگو دربرمی‌گیرند. اطلاعات به معنی ارتباط دانش درباره یک رویداد یا شرایط یا گسترش دانش بدست آمده از مشاهده، مطالعه یا آموزش است. کلماتی مثل دانش، حقیقت، داده، اخبار و پیام، نیز به عنوان هم‌معنی یا نزدیک به معنی اطلاعات استفاده می‌شود. در علوم اطلاعات [۲۹]، مفهوم اطلاعات به روش‌های مختلفی بیان شده است: اینورسن^۲ معتقد است که اطلاعات چیزی است که از انتقال یک ساختار شناختی توسط ذهن، مدل وضعیت دانش دریافت‌کننده و در فرمی از نشانه‌ها است، نتیجه می‌شود [۲۴]. به عبارت دیگر اطلاعات ساختاری است که وقتی فراهم شد، ممکن است روی وضعیت دانش فرد دریافت‌کننده تاثیر بگذارد و وضعیت وی را تغییر دهد. دروین^۳ و میلان^۴ معتقدند که اطلاعات چیزی است که توسط انسان ایجاد شده است [۱۶]. ترور^۵، اطلاعات را یک منبع کلیدی می‌داند که می‌تواند در جامعه تغییر و پیشرفت ایجاد کند [۱۵]. از دیدگاه شناختی علم اطلاعات، بلکین^۶ معتقد است که اطلاعات در رابطه با یک

¹ Information

² Ingwersen

³ Dervin

⁴ Milan

⁵ Turnor

⁶ Belkin

متن، ساختاری است که بر ساختار سطحی آن متن تاکید می‌کند (مثل زبان). الگوی وی بر این فرضیه استوار است که نیاز به اطلاعات نتیجه "ناهمگونی در وضعیت علمی" یک فرد است. زیرا افراد نمی‌توانند آنچه را که نمی‌دانند یا جایش در ذهنشان خالی است به آسانی بیان کنند. همچنین وی اعتقاد دارد سؤالاتی که طبق درخواست افراد به نظام‌های اطلاعاتی ارجاع می‌شود، نمی‌تواند گویای نیازهای اطلاعاتی واقعی آن‌ها باشد. برای حل این مشکل، وی عقیده دارد که پی بردن به موقعیت کاربر و فاصله (آنچه بلکین ناهمگونی می‌نامد) بسیار حایز اهمیت است [۱۱].

از آنجایی که در پژوهش‌های علوم اطلاعات، قسمت مهم و قابل توجه، رفتار مرتبط با اطلاعات است که شامل جستجوی اطلاعات مورد نیاز، رفتار جستجوی اطلاعات و مدل‌های جستجوی اطلاعات است [۴۱]، در بخش بعدی تعاریف و مدل‌های آن‌ها بیان شده است.

۲-۱-۱- جستجوی اطلاعات^۱ و رفتار جستجوی اطلاعات^۲

از نظر وبرلی^۳ جستجوی اطلاعات یک فعالیت پایه‌ای و اساسی است که توسط همه‌ی افراد تحقق می‌یابد [۵۱]. رفتار اطلاعاتی، کلیه‌ی رفتارهای اطلاعاتی انسان در رابطه با منابع و کانال‌های اطلاعاتی است که شامل جستجوی فعال و غیرفعال اطلاعات و استفاده از اطلاعات است [۲۹]. در واقع، رفتار جستجوی اطلاعات یک واژه‌ی گسترده است که شامل راه‌هایی است که افراد نیازهای اطلاعاتی‌شان را بیان می‌کنند؛ اطلاعات را جستجو، ارزیابی، انتخاب و از آن استفاده می‌کنند [۸]. [۴۱]. بنابراین رفتار جستجوی اطلاعاتی، هدف جستجوی یک نیاز برای ارضای برخی اهداف دیگر است [۲۹]. از نظر ویلسون^۴، جستجوی اطلاعاتی توصیف روش‌هایی است که افراد اطلاعات را جستجو، بررسی، ارزیابی و استفاده می‌کنند. در مسیر جستجوی اطلاعات جدید، فرد ممکن است با افراد مختلف، ابزارهای آنالوگ، کتابخانه‌ها، مراکز اطلاعاتی و سیستم‌های اطلاعاتی براساس کامپیوتر تعامل داشته باشد [۵۲]. از یک دیدگاه دیگر فرآیند کسب اطلاعات، استفاده و به کار بردن اطلاعات، به عنوان رفتار جستجوی اطلاعاتی شناخته شده است. واضح است که رفتارهای اطلاعاتی در محیط‌های

¹ Information Seeking

² Information Seeking Behaviour

³ Wiberley

⁴ Wilson

علمی و دانشگاهی، به دلیل نیاز پژوهشگران و دانشجویان به اطلاعات درست و به روز برای پژوهش‌هایشان، نقش بسیار مهمتری دارد [۴۱]. براساس پژوهش‌هایی که درباره الگوهای جستجوی اطلاعاتی از پژوهشگران در کتابخانه‌های دانشگاهی انجام شده است، نتایج نشان می‌دهد که حتی بین الگوهای مورد استفاده از اطلاعات میان پژوهشگران رشته‌های مختلف تفاوت وجود دارد [۴۴] و عوامل گوناگونی از جمله سهولت و سرعت دسترسی، آشنایی با منابع، سطح و موضوع پژوهش بر روی رفتار جستجو تاثیر گذارند [۸]. در بخش‌های بعدی انواع مدل‌های جستجوی اطلاعات برای رفع نیازهای کاربران بررسی می‌شود.

۲-۲-۱- مدل‌های جستجوی اطلاعات^۱

در این بخش مدل‌های اطلاعاتی مورد بررسی قرار می‌گیرند.

۱-۲-۲-۱- مدل جستجوی اطلاعات ویلسون و کیرکلاس^۲

این مدل رابطه بین مفاهیم کاربر، نیاز، کاربردها و رفتار کاربر را نشان می‌دهد. مدل، فهم نیاز کاربر در زمینه محیط وی را پیشنهاد می‌کند، به این معنی که در یک محیط یا رویداد (مثلا دروس دانشگاه) نیاز اطلاعاتی کاربر را درک می‌کند. نیاز درک شده منجر به جستجوی اطلاعات توسط کاربر می‌شود. منابع اطلاعاتی جهت جستجو شامل سیستم‌های اطلاعاتی (کتابخانه‌های دانشگاه و کتابخانه‌های عمومی)، منابع انسانی (خبره‌ها، اساتید، همکاران)، و دیگر منابع (کتابخانه‌های شخصی، رسانه) هستند [۲۹].

۲-۲-۲-۱- مدل فرآیند جستجوی اطلاعات بلکین^۳

بلکین دیدگاه دیگری از فرآیند جستجوی اطلاعات را تحت عنوان استراتژی‌های جستجوی اطلاعات ارائه کرد. مجموعه فعالیت‌ها در این فرآیند شامل موارد زیر هستند:

¹ Model Of Information Seeking

² Krieklas

³ Belkin

- مرور^۱: اسکن و جستجوی یک منبع
 - یادگیری^۲: گسترش دانش بر اساس هدف، مسئله، سیستم یا منابع در دسترس از طریق انتخاب
 - تشخیص^۳: تشخیص آیتم‌های مربوطه (از طریق سیستم یا روابط شناختی)
 - ابراطلاعات^۴: تعامل با آیتم‌هایی که مرز فعالیت‌ها را نگاشت می‌کند
- این مدل، خطی یا شبیه به یک فرایند جریان آبخاری نیست. بلکه تاکید دارد ویژگی غیرخطی بودن که وی پیشنهاد داده است از "حرکت ظریف"^۵ بین فعالیت‌ها پشتیبانی می‌کند [۲۹].

۳-۲-۱- مدل شش مهارتی ایزنبرگ^۶ و برکویتز^۷

ایزنبرگ و برکویتز یک دیدگاه کلی برای حل مسئله اطلاعات شامل شش مرحله منطقی را پیشنهاد دادند [۲۹]:

- تعریف فعالیت^۸
- استراتژی‌های جستجوی رفتار^۹
- مکان‌یابی و دسترسی^{۱۰}
- استفاده از اطلاعات^{۱۱}
- ترکیب^{۱۲}
- ارزیابی^{۱۳}

¹ Browsing

² Learning

³ Recognition

⁴ Meta information

⁵ Graceful movement

⁶ Eisenberg

⁷ Berkowitz

⁸ Task Definition

⁹ Information Seeking Strategies

¹⁰ Locating and access

¹¹ Use of information

¹² Synthesis

¹³ Evaluation

۴-۲-۱-مدل جستجوی اطلاعات الیس^۱

الیس، یک مدل کلی از رفتارهای جستجوی اطلاعات را براساس مطالعاتی که از الگوهای جستجوی اطلاعات دانشمندان علوم اجتماعی، فیزیکدانان، شیمیدانان، مهندسان و پژوهشگران در صنعت کرده بود، پیشنهاد داد[۲۹]. این مدل شامل مراحل زیر می‌باشد:

- راه‌اندازی یا شروع^۲
- زنجیر کردن^۳
- مرور
- تمایز^۴
- نظارت^۵
- استخراج^۶

۳-۱- جستجوی اطلاعات در اینترنت

امروزه، شبکه جهانی وب، منبع عظیمی از اطلاعات است که این اطلاعات با محتوای تحت وب توسط میلیون‌ها صفحه در دسترس عموم مردم قرار دارد[۲۰][۲۷]. افراد از طریق موتورهای جستجو اطلاعات مورد نیاز خود را از این منابع اطلاعاتی دریافت می‌کنند. در طول دهه گذشته، موتورهای جستجو به ابزار اصلی پیدا کردن اطلاعات برای بسیاری از مردم تبدیل شده‌اند[۲۱]. موتورهای جستجو به طور چشمگیری بخش مهمی از فعالیت کاربران اینترنتی هستند. به طورمثال، گزارش‌ها نشان می‌دهد که روزانه حدود ۶۰ میلیون از بزرگسالان آمریکایی از موتورهای جستجو استفاده می‌کنند[۴۵]. همچنین در

¹ Ellis

² Starting

³ Chaining

⁴ Differentiating

⁵ Monitoring

⁶ Extracting

تحقیقی که توسط نیلسن مدیا^۱ انجام شده است، گزارش شد که بیش از ۸۰ درصد از جستجوگران وب از موتورهای جستجوی وب برای یافتن اطلاعات و خدمات آنلاین استفاده می‌کنند[۳۷]. بنابراین موتورهای جستجو ابزار مهمی در کسب اطلاعات عمومی افراد است.

از طرفی در دهه‌های اخیر شاهد خلاقیت و نوآوری در ارتباط علمی میان دانشمندان و پژوهشگران در سراسر جهان هستیم که این امر تا حد زیادی به توسعه ارتباطات و انتشارات الکترونیکی قابل استناد کمک نموده است[۸]. انتشار محتواهای علمی و دستاوردهای پژوهشی به صورت الکترونیکی به ویژه بر روی اینترنت، روش توزیع اطلاعات علمی در دنیا را تغییر داده است[۲۲]. با توسعه انتشارات الکترونیکی و دیجیتالی شدن نه تنها توسعه کتابخانه‌ها بهبود یافته است بلکه آن‌ها را از منابع نگهدارندگان اطلاعات به منابع دسترسی اطلاعات تغییر داده است[۱۹].

طبق پژوهشی که اخیراً در سال ۲۰۱۵، توسط آهیاوز^۲ و آنی^۳ صورت گرفت[۸]، الگوهای رفتاری جستجوی اطلاعات ۱۲۰ تن از کارمندان دانشگاهی در دانشگاه علوم و فناوری ریورس^۴ بررسی شد، نتایج نشان می‌دهد که اینترنت بالاترین رتبه‌بندی را با امتیاز ۳۵۷ (۳۱٫۶ درصد) داشته است و تماس‌های اشخاص با ۲۴ درصد، کارگاه و کنفرانس‌ها با ۲۳ درصد و کتابخانه‌ی دانشگاه با ۲۰ درصد، در رتبه‌های بعدی قرار می‌گیرند.

حال با توجه به اینکه اینترنت یک منبع اطلاعاتی غنی برای تمامی کاربران و علی‌الخصوص کاربران دانشگاهی است، توجه به کیفیت موتورهای جستجو از اهمیت ویژه‌ای برخوردار است. دو شاخص کلیدی که کیفیت موتورهای جستجو را به طور کلی مشخص می‌کند[۳۰][۴۵] عبارتند از: (۱) کیفیت نتایج برگشت داده شده (۲) سرعت نتایجی که برگشت داده شده. برای ارتقای شاخص اول، فهم کاربران و نیازهای آنان راه حل اصلی است[۲۷][۵۰]. به همین دلیل، پژوهشگران به بررسی لاگ‌های جستار^۵ جستجو می‌پردازند. بررسی لاگ‌های جستار به محققان کمک می‌کند تا درک

¹ Nielsen Media

² Ahiauz

³ Ani

⁴ Rivers State University of Science and Technology (RSUST)

⁵ Query

بهتری از آنچه کاربران به دنبالش هستند، داشته باشند و با فهم نیاز کاربران بتوانند فرآیند جستجو را بهبود دهند. تمرکز اصلی این پژوهش نیز بر روی فهم کاربران و نیازهای آنان از طریق بررسی لاگ‌های جستجوی سامانه‌ی گنج می‌باشد.

۱-۴- تحلیل لاگ^۱

لاگ یک رکورد الکترونیکی از تعاملاتی است که بین سیستم و کاربران سیستم می‌افتد. فایل‌های لاگ می‌تواند از منابع مختلفی از جمله از وب سایت‌ها، کامپیوتر کاربر، بلاگ‌ها، روزنامه‌های آنلاین و غیره استخراج شوند [۲۴]. داده‌های لاگ، داده‌هایی غنی همراه با جزئیات فردی هستند [۳۵]. تحلیل لاگ، فرآیند تبدیل داده‌های خام به اطلاعاتی برای حل مسائل است [۷]. در واقع تحلیل لاگ، روش‌شناسی جهت مطالعه سیستم‌های آنلاین و کاربران این سیستم‌هاست [۲۴]. پیترز^۲، تحلیل لاگ را مطالعه‌ی تعاملات الکترونیکی ضبط شده بین سیستم‌های بازیابی اطلاعات آنلاین و افرادی که در این سیستم‌ها جستجو می‌کنند، تعریف کرد [۴۲]. مباحث لاگ‌ها و تحلیل لاگ‌ها در کل روشی جهت جمع‌آوری داده و بررسی کارایی سیستم و تحلیل رفتار کاربر است که از سال ۱۹۶۷ تاکنون استفاده می‌شود [۳۳].

تحلیل لاگ طبقه‌بندی گسترده‌ای را شامل می‌شود از جمله تحلیل وب لاگ‌ها (تحلیل لاگ‌های سیستم وب)، تحلیل بلاگ و تحلیل لاگ جستجو (تحلیل لاگ‌های موتور جستجو) [۲۴].

لاگ کاوی جستار^۳، یک شاخه‌ی عمومی‌تر از لاگ کاوی^۴ و تحلیل لاگ است [۲۵]. در عمل، لاگ کاوی جستار، با بهره‌گیری از یکسری تکنیک‌ها که هدفشان کشف الگوهای مورد علاقه از لاگ‌های جستارهای موتورهای جستوی وب است، تلاش می‌کند کارایی سرویس‌های آنلاین افزایش یابد [۴۵]. فرمت لاگ جستار معمولاً به صورت یک رکورد است که شامل ویژگی‌هایی از جمله شناسه کاربر، زمان، رده‌بندی نتایج کلیک شده و رشته URL کلیک شده است [۴۵].

¹ Log analysis

² Peters

³ Query Log Mining

⁴ Log mining

۱-۴-۱- مزایا و معایب جمع‌آوری و تحلیل داده‌های لاگ جستار

جمع‌آوری و تحلیل لاگ‌های جستار شامل مزایایی است از جمله: بهبود جستجوی عمومی و جستجوی شخصی، امکان ثبت تاریخچه‌ی جستجوی افراد در سیستم و ارزانتر بودن هزینه جمع‌آوری لاگ‌ها نسبت به جمع‌آوری داده‌های اصلی [۲۴]

[۳۵]. در کنار مزایای ذکر شده برای تحلیل و جمع‌آوری لاگ‌های جستار، خطراتی نیز وجود دارد از جمله اینکه داده‌ها ممکن است توسط موسسه جمع‌آوری کننده مورد سوء استفاده قرار بگیرند و یا توسط افراد با نیت مخرب به بیرون درز پیدا کنند. همچنین ممکن است توسط دولت‌ها استفاده شوند. عیب بزرگ داده‌های لاگ این است که معمولاً داده‌های جامع و کاملی نیستند، زیرا با توجه به سوالات یک پژوهش خاص جمع‌آوری نشده‌اند. معمولاً برای بررسی محدوده خاصی در حوزه لاگ کاوی که برخی از پژوهشگران تمایل به مطالعه آن دارند، داده‌های لاگ داده‌های مناسبی نیستند. نکته دیگر این است که در زمان جمع‌آوری این داده‌ها اگر کاربران سیستم از جمع‌آوری لاگ‌هایشان آگاه باشند ممکن است رفتار خود را تغییر دهند و رفتارهای آن‌ها کاملاً طبیعی نیست [۳۵].

۲-۴-۱- محدودیت‌های تحلیل لاگ‌ها

در سال ۱۹۹۳، کورت^۱ محدودیت‌هایی که بر سر راه تحلیل لاگ‌ها وجود دارد را بررسی کرد [۳۰]. از نظر وی چهار عنصر، چارچوب تحلیل لاگ‌ها را شکل می‌دهند: سیستم‌های آنلاین، کاربر و فرآیند جستجو، تحلیل داده‌های لاگ‌ها و مسائل اخلاقی و قانونی. عوامل سیستمی که مانع جمع‌آوری داده‌های لاگ‌ها می‌شوند عبارتند از عوامل مکانیکی، موقتی (زمانی)، فضایی^۲ و عوامل داخلی. مهمترین محدودیت‌های مکانیکی، نیازمندی‌های برنامه‌نویسی و سایز داده‌ها هستند. زیرا تصمیم‌گیری برای مدیران سیستم را دشوار می‌کند و آن‌ها را مجبور به متعادل‌سازی رفع نیاز برای ذخیره داده‌های نشست^۴ و نیازهای دیگر سیستم می‌کند. عوامل موقتی (زمانی) و فضایی که مانع از مطالعه یک سیستم آنلاین میشوند وقتی مطالعه دو یا بیش از دو سیستم مطرح است به صورت نمادین افزایش پیدا می‌کنند. عوامل داخلی هم زمانی که لاگ‌های دو سیستم

^۱ Kurht

^۲ Temporal

^۳ Spatial

^۴ Session

قرار است مقایسه شوند، ایجاد دردسر می‌نماید. زیرا هر سیستمی نحوه ذخیره‌سازی لاگ‌های منحصر به فرد دارد و امکان مقایسه را مشکل‌ساز می‌کند.

عوامل کاربر و فرآیند جستجو که تحلیل لاگ‌ها را مشکل می‌سازد، مسئله عدم امکان ایزوله‌سازی کاربران سیستم‌های آنلاین است تا بتوان الگوهای رفتاری‌شان را به طور دقیق استخراج کرد. عوامل تحلیل داده‌ها شامل تشخیص نشست‌های جستجوی افراد می‌شود [۱۴]، زیرا معمولاً نیاز به بررسی خط به خط لاگ‌ها جهت تشخیص اولین خط هر نشست جدید است. انتخاب نقاط ابتدایی و انتهایی نشست‌های جستجو دشوار است و بر روی خروجی تحلیل تاثیرگذار است. فاکتورهای قانونی و اخلاقی نیز به نوبه خود موانعی را در پژوهش‌های مربوط به لاگ‌ها ایجاد می‌کنند [۳۰]. مسئله‌ی حفظ حریم خصوصی کاربران به عنوان مانع عمده برای نظارت گسترده بر سیستم‌های بازیابی اطلاعات است. پژوهشگران، در زمان مطالعه و جمع‌آوری داده‌های نشست جستجوی سیستم‌های آنلاین، از لحاظ قانونی و اخلاقی موظف به حفظ حقوق حریم خصوصی کاربران در فعالیت‌های جستجوی آنلاین‌شان هستند.

۳-۴-۱- تقسیم‌بندی جستارها

تقسیم‌بندی‌های مختلفی برای جستارهای جستجو در موتورهای جستجو ارائه شده است. در سال ۲۰۰۲، بردر^۱، جستارها را در سه دسته: هدایتی^۲، اطلاعاتی^۳ و تراکشی^۴ تقسیم‌بندی نمود [۱۳]. در سال ۲۰۰۴، رز^۵ و لوینسون^۶ تقسیم‌بندی با جزییات بیشتری را برای تقسیم‌بندی جستارها براساس اهداف جستجو ارائه کردند [۴۳]:

الف) هدایتی: در این نوع جستارها، کاربر اطلاعات یک وب سایت یا یک صفحه‌ای را می‌داند که قرار است وجود داشته باشد. جستارهای ناوبری یا هدایتی معمولاً حاوی قطعاتی از آدرس‌ها و یا اسامی سازمان‌ها هستند. معمولاً، کاربر فقط بر روی یک نتیجه کلیک می‌کند که آن نتیجه وی را به طور مستقیم به صفحه مورد نظر هدایت می‌کند.

¹ Border

² Navigational

³ Informational

⁴ Transactional

⁵ Rose

⁶ Levinson

ب) اطلاعاتی: این جستارها شبیه به جستارهایی است که به طور سنتی در بازیابی اطلاعات مورد مطالعه هستند. برای مثال، کاربر می‌خواهد روی یک موضوع خاص، چه به طور گسترده (مثلا "تاریخ ما") و یا محدود (مثلا "تغذیه ویژه برای مراقبت از زخم") اطلاعاتی بدست آورد. در این حالت کاربران اغلب چندین لینک را به عنوان نتیجه دنبال می‌کنند. این نوع جستارها شامل موارد زیر است:

- بسته^۱: جستارهایی که به دنبال یک پاسخ، بسته شده است.
- باز^۲: جستارهایی که انتهای پاسخ آن باز است و یا پاسخ‌هایی با عمق نامحدود دارد.
- غیرمستقیم^۳: جستارهایی که هدفشان همه چیز یا یک چیز درباره یک موضوع خاص است.
- توصیه^۴: جستارهایی که توصیه یا دستوراتی را برای تکمیل یک فعالیت جستجو می‌کنند.
- مکان‌یابی^۵: جستارهایی که برای شناسایی یک مکان و یا دریافت خدمات تلاش می‌کنند.
- فهرست^۶: جستارهایی که صفحات خوب روی یک موضوع را جستجو می‌کنند.
- ج) منابع^۷: هدف این جستارها، منابع هستند و نه صفحات وب.
- دانلود^۸: جستارهایی که هدف آن‌ها دانلود یک منبع مفید است.
- سرگرمی^۹: جستجو برای صفحاتی که می‌تواند سرگرمی را فراهم کند.
- تعامل^{۱۰}: جستارهایی که صفحاتی را که نیاز به تعامل دارند جستجو می‌کنند برای مثال، یک نقشه و یا سرویس - های آب و هوا.
- Obtain: اسنادی مانند فرم‌های مالیاتی یا اسناد دولتی را جستجو می‌کنند.

¹ closed

² open

³ undirected

⁴ advice

⁵ locate

⁶ list

⁷ resource

⁸ download

⁹ Entertainment

¹⁰ Interact

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۱۲

از دیدگاه دیگری [۲۱]، جستارها را مطابق با موضوعی که نزدیک به نیت جستجو می‌باشد نیز در دسته‌های زیر

تقسیم‌بندی می‌کنند:

- سرگرمی
- پرنوگرافی
- کسب و کار، سفر، استخدام
- کامپیوتر
- علم و پزشکی
- افراد، مکان، خدمات
- اجتماعی و مذهبی
- آموزش، علایق انسانی
- هنر
- دولت (حکومت)
- ناشناخته و دیگر

از دیدگاه [۱۲] جستارها براساس طول‌شان به دو دسته کوتاه و بلند تقسیم می‌شوند. اکثر جستارها در لاگ جستجو، کوتاه

هستند. جستارهایی با طول کمتر از ۴، کوتاه و جستارهای بلند طولشان بین ۵ تا ۱۲ است. جستارهای بلند خود نیز به پنج

دسته تقسیم‌بندی می‌شوند:

- جستارهای سوالی^۱: این نوع جستارها شامل یکی از کلمات زیر هستند: چه چیزی، چیست، چه کسی، کجا، چه زمانی، چرا، چگونه، کدام یک، چه کسی، متعلق به چه کسی، آیا، هست، وجود دارد.
- جستارهای اپراتورها^۱: این مدل از جستارها شامل حداقل یکی از اپراتورهای بولین AND، OR و NOT هستند.

¹ Questions (QE)

- جستارهای ترکیبی^۲: ترکیبی از جستارهای کوتاه هستند.
- جستارهای غیر ترکیبی^۳: جستارهایی که نمی‌توانند به عنوان یک ترکیب از انواع جستارهای کوتاه تعریف شوند. این نوع جستارها شامل دو نوع: عبارت اسمی^۴ و عبارت فعلی^۵ هستند.

۴-۴-۱- سوالات متداول در پژوهش‌های بررسی لاگ‌ها

پس از شناخت مفاهیم اولیه در ارتباط با اطلاعات و روش‌های جستجوی آن و بررسی محدودیت‌ها و چالش‌های تحلیل لاگ، باید سوالاتی که در هر پروژه‌ی پژوهشی که در ارتباط با لاگ‌ها صورت می‌گیرد، را بررسی کنیم. تا بدانیم در انجام این سبک پروژه‌ها چه نوع دغدغه‌های مشترکی وجود دارد و باید چه سوالاتی را پاسخگو باشیم. سوالات متداول در این حوزه پژوهشی عبارتند از [۲۴]:

۱. کدام داده‌ها تجزیه و تحلیل می‌شوند؟ پژوهشگر به وضوح باید شیوه و فرمت اطلاعات لاگ ثبت شده را بیان کند.
۲. این اطلاعات چگونه تعریف شده‌اند؟ پژوهشگر باید هر کدام از معیارهایی که جهت ثبت در نظر گرفته شده‌اند را به وضوح تعریف کند به طوری که اجازه جابجایی آن در سیستم‌های دیگر و با دیگر کاربران فراهم شود.
۳. جمعیتی که پژوهشگر داده‌ها را از آن استخراج می‌کند چه کسانی هستند؟ پژوهشگر باید از کاربران و سیستمی، که داده‌های آن‌ها را ردیابی و ثبت می‌کند آگاه باشد. این امر گاهی اوقات بسیار دشوار است مگر این که سیستم به برخی از انواع لاگین‌ها نیاز داشته باشد و این پروفایل‌ها در دسترس باشند. در غیاب این پروفایل‌ها پژوهشگر باید به بررسی‌های جمعیتی و مطالعات مربوط به جمعیت کاربران یا جمعیت وب عمومی اتکا کند.
۴. پژوهشگر در چه زمینه‌ای داده‌ها را تجزیه و تحلیل کرده است؟ مهم است که پژوهشگر به وضوح فاکتورهای محیطی، موقعیتی و محتوایی را که تحت آن‌ها داده‌ها را جمع‌آوری کرده بیان کند.

¹ Operators (OP)

² Composite (CO)

³ Non-Composite

⁴ Noun-phrases (NC_NO)

⁵ Verb-phrases (NC_VE)

۵. چه مرزهایی برای تحلیل در نظر گرفته شده است؟ پژوهش با استفاده از ردیابی داده‌ها از لاگ‌ها گول‌زننده است و پژوهشگر باید در استفاده از سوالات پژوهش و یافته‌های آن مراقب باشد. برای مثال با داده‌های لاگ، آیا کاربر بر روی یک لینک کلیک کرده است یا خیر و یا چرا کاربر بر روی یک لینک کلیک کرده است.
۶. هدف از استنتاج چیست؟ پژوهشگر بایستی بوضوح روابط بین معیارهای مجزا در ردیابی اطلاعات را برای استنتاج بیان کند. این شرح و استنتاج می‌تواند در هر سطحی از جزییات (فردی، مجموعه‌ای از افراد، سازمان و غیره) باشد.

۵-۱- نتیجه‌گیری

در این بخش مفاهیم کلی اطلاعات، جستجوی اطلاعات و انواع روش‌های آن مورد بررسی قرار گرفت. همچنین تحلیل لاگ به عنوان یکی از منابع غنی ذخیره‌سازی رفتار کاربران مورد بررسی قرار گرفت و مزایا، معایب و محدودیت‌های تحلیل لاگ بیان شد.

فصل دوم

کارهای پیشین

۱-۲- مقدمه

تحلیل لاگ سامانه‌های اطلاعاتی روشی است که از ۱۹۶۰ تاکنون جهت ارزیابی سامانه‌ها و نیز مطالعه‌ی رفتار کاربران مورد استفاده بوده است [۷]. با توجه به مطالعات انجام شده در این حوزه، پژوهش‌های انجام شده تاکنون را به دو قسمت تقسیم نمودیم: پژوهش‌های بخش اول بر پژوهش‌های انجام شده بر روی انواع موتورهای جستجو و بررسی لاگ‌ها، جستارهای مطرح بر روی آن‌ها و بررسی روش‌های خوشه‌بندی جستارها تاکید دارند. بخش دوم، پژوهش‌هایی هستند که صرفاً بر روی موتورهای جستجوی پایگاه‌های علمی مانند ScienceDirect تمرکز دارند.

۲-۲- تحلیل لاگ موتورهای جستجوی پایگاه‌های غیر علمی

مارچيونینی^۱ [۳۲] معتقد است که فرآیند جستجوی اطلاعات با چندین عامل از قبیل ویژگی‌های فردی، نوع جستجو، ویژگی سامانه‌های اطلاعاتی، حوزه‌های دانش، محیط اطلاعاتی، تجربه و دانش پیشین در جستجو و نتایج جستجو (بازخورد از سامانه) در ارتباط است. براساس این رویکرد، کاربر مرکز اصلی یک سامانه‌ی اطلاعاتی است. از این‌رو پژوهش‌های فراوانی با استفاده از تحلیل لاگ موتورهای جستجوی وب منتشر شده است که نشان دهنده‌ی چگونگی رفتار جستجوی کاربران وب هستند.

در سال ۲۰۰۰، جانسن^۲ و همکارانش [۲۶] گزارش کردند که ۶۷ درصد از کاربران فقط یک جستجو را در موتور جستجو انجام داده‌اند و ۱۹ درصد آن‌ها تنها یک تغییر در جستجوی خود ایجاد کرده‌اند.

در سال ۲۰۰۰، مورزی^۳ و همکارانش یک الگوریتم جدید برای تقسیم‌بندی خودکار کاربران وب براساس الگوهای دسترسی‌شان ارائه دادند که منجر به بهبود سازماندهی اسناد وب برای راحتی هدایت کاربران به سمت نیازهایشان شد [۳۴]. این الگوریتم از خوشه‌بندی کاربران براساس لاگ‌های دسترسی کاربران بهره می‌جوید و کاربران را به گونه‌ای خوشه‌بندی می‌کند که کاربران داخل یک خوشه شبیه به همدیگرند و با خوشه‌های دیگر متفاوت‌اند. در واقع، کاربرانی که

¹ Marchionini

² Jansen

³ Morzy

در دسترسی به منابع اطلاعاتی مسیرهای مشابهی را دنبال کرده‌اند در یک خوشه قرار می‌گیرند. الگوریتم پیشنهادی از الگوهای تکراری برای تولید مدل خوشه‌بندی و محتوای خوشه استفاده می‌کند. الگوریتم به صورت تکراری خوشه‌های کوچکتر و مشابه را ادغام می‌کند تا وقتی که تعداد خوشه‌های درخواست شده تمام شود. از معیار شباهت داخل خوشه‌ای^۱ جهت ادغام خوشه‌ها براساس هم‌رخدادی استفاده می‌شود. یکی از ویژگی‌های مهم این الگوریتم این است که نه تنها کاربران وب را به خوشه‌ها تقسیم می‌کنند، بلکه یک مدل طبقه‌بندی نیز ارائه می‌شود که می‌تواند برای طبقه‌بندی کاربران وب در آینده استفاده شود. از آنجایی که مدل با یک مجموعه از الگوهای تکراری مسیرهای دسترسی شکل گرفته، طبقه‌بندی یک مسیر دسترسی کاربر جدید به سادگی امکان‌پذیر است که آیا وی همان مسیرهای دسترسی را طی می‌کند و متعلق به خوشه‌های از قبل مشخص شده است یا مسیر دسترسی کاربر جدید شامل الگوهایی از خوشه‌های متفاوت است.

در سال ۲۰۰۱، زی^۲ و همکارانش روشی جهت مقایسه و تحلیل لاگ‌های جستار کاربران ارائه دادند [۵۴] که هدف آن‌ها پاسخ به این سوال بود که آیا جستارهای کاربران در طول زمان تغییر می‌کند؟ و اگر تغییر وجود دارد بر چه اساسی است؟ و آیا امکان پیش‌بینی جستارهای آینده‌ی کاربران، بر اساس مشاهده‌ی جستارهای گذشته‌ی آن‌ها وجود دارد؟ جهت پاسخ به این سوالات، پژوهشگران این پژوهش از روش‌های آماری جهت اندازه‌گیری شباهت لاگ‌ها استفاده کردند. داده‌های لاگ این پژوهش از سرویس جستجوی AOL شرکت آنلاین آمریکایی (<http://www.aol.com>) از ماه می ۲۰۰۱ تا ماه آوریل ۲۰۰۲ جمع‌آوری شد. نتایج نشان داد که تعداد اندکی از جستارها (تقریباً ۱۰۰۰۰ تا) سهم زیادی از جستارها را شامل می‌شوند که حدوداً بیش از ۸۰٪ کل جستارهاست. در واقع اگر این درصد اندک از جستارها قابل پیش‌بینی باشند و نتایج جستجو برای آن‌ها بهینه شود، کل سیستم می‌تواند برای ۸۰٪ ارائه دهنده‌های خدمات بهینه شود. همچنین نتایج نشان داد که تعداد ۱۰۰۰ جستار نخست از لحاظ بیشترین تکرار به طور متوسط ۸۸٫۱٪ در هر ماه با ماه بعد، ۸۲٪ فصل به فصل و ۷۸٫۵٪ هر شش ماه مشابه هستند. به علاوه مقدار انحراف معیار پایین نشان می‌دهد که جستارها در لاگ‌ها پایدارند. نتایج تحلیل لاگ‌ها گزارش می‌دهد که جستارهایی که در طول یک سال دوام آورده‌اند فقط حدود ۲۰ الی ۳۰٪

^۱ inter-cluster

^۲ Xi

همه‌ی جستارهای منحصر به فرد هستند. نتایج نشان می‌دهد که هرچه یک جستار در طول زمان بیشتر دوام بیاورد احتمال دوامش در آینده نیز بیشتر است. لذا نیازهای اطلاعاتی کاربران در طول زمان بسیار ثابت است و فقط تنها ۲۰٪ در طول زمان تغییر می‌کنند. لذا می‌توان درخواست‌های آینده‌ی کاربران را براساس لاگ‌های درخواست‌های گذشته‌ی آنان با ضریب اطمینان ۸۰٪ پیش‌بینی نمود.

در سال ۲۰۰۲، ون^۱ و همکارانش یک روش خوشه‌بندی جدید برای جستارها را ارائه کردند که در آن از لاگ‌های کاربر استفاده می‌کنند [۴۹]. این روش امکان تشخیص اسنادی که کاربران برای یک جستار انتخاب کرده‌اند را فراهم می‌کند. در واقع شباهت بین دو جستار براساس اسنادی که کاربران برای آن انتخاب کرده‌اند، مشخص می‌شود. خوشه‌بندی جستارها فرآیندی جهت کشف سوالات اغلب پرسیده شده^۲ یا موضوعات معروف روی یک موتور جستجو است. روش‌های مختلفی جهت خوشه‌بندی جستارها استفاده می‌شود [۴۹]:

استفاده از کلمات کلیدی: در این روش اسناد براساس کلمات کلیدی موجود در آن‌ها خوشه‌بندی می‌شوند. در این روش به طور کلی یک سند به صورت یک بردار در یک فضای برداری که توسط همه‌ی کلمات کلیدی آن شکل یافته نمایش داده می‌شود. پژوهش‌ها در این روش معمولاً بر روی انتخاب تابع شباهت و الگوریتم‌هایی جهت فرآیند خوشه‌بندی تاکید دارند. به طور معمول، از معیار شباهت cosine، تشابه ژاکارد^۳ یا Dicesimilarity، و دربرخی موارد ویرایش فاصله برای تابع شباهت استفاده می‌شود. الگوریتم‌های خوشه‌بندی نیز شامل الگوریتم خوشه‌بندی زنجیره‌ای سلسله مراتبی (HAC) و k-means هستند.

^۱ WEN

^۲ Frequently asked questions (FAQ)

^۳ Jaccard

استفاده از ابر لینک‌ها:^۱ بخاطر محدودیت‌های کلمات کلیدی، پژوهشگران سعی دارند از معیارهای دیگری جهت خوشه‌بندی استفاده کنند که یکی از آن‌ها لینک بین اسناد است. فرض این است که ابرلینک‌ها اسناد مشابه را به یکدیگر متصل می‌کنند. این ایده در سیستم‌های بازیابی اطلاعات از جمله گوگل استفاده می‌شود.

استفاده از Cross-reference بین جستارها و اسناد: در روش استفاده از ابرلینک، فضای اسناد و فضای جستار از هم جدا هستند. ایده‌ی استفاده از cross-references، کمک می‌کند تا به هر رابطه‌ی ایجاد شده بین یک جستار و یک سند معنی داده شود. دیدگاه استفاده از cross-references بر شباهت بین اسناد که می‌توانند از طریق اسناد به جستارها منتقل شد و برعکس تاکید دارد. در واقع اگر تعداد زیادی از افراد یک مجموعه از اسناد را برای جستارهای یکسانی انتخاب کنند، این یک نشانه است که این اسناد مشابه‌اند. این فرضیه در موتور جستجوی CiteSeer (<http://citeseer.nj.nec.com/>) استفاده می‌شود. یعنی یک مرتبه که یک سند شناسایی می‌شود، سیستم مجموعه‌ای از اسنادی را که دیگران به آن سند دسترسی داشته‌اند را نمایش می‌دهد. موتور آمازون (<http://amazon.com>) نیز از همین فرضیه جهت نمایش کتاب‌های مشابه به کاربران مشابه استفاده می‌کند.

روشی که ون^۲ و همکارانش [۴۹] استفاده کردند براساس دو معیار: جستار و cross-references است که این دو با اصول زیر مشخص شدند: اصل ۱ (استفاده از محتوای جستار): اگر دو جستار شامل اصطلاحات یکسان یا مشابه باشند، آن‌ها نیازهای اطلاعاتی مشابهی را مشخص می‌کنند. همچنین واضح است که جستارهای طولانی‌تر، قابل اعتمادتر هستند. البته چون این اصل به تنهایی کافی نیست، بنابراین معیار دوم برای تکمیل استفاده می‌شود. اصل ۲ (استفاده از کلیک‌های سند): دو جستار مشابه‌اند اگر آن‌ها به انتخاب سند مشابهی منجر شوند. اصل ۱ به تنهایی جهت خوشه‌بندی جستارهای مرتبط با معنا (معنا شناسی) کافی نیست. چون هدف اصلی آن کلمات کلیدی است و معنا جایگاهی ندارد. مشخص است که کلمات کلیدی یکسان همواره نیاز اطلاعات یکسانی را نشان نمی‌دهند. به طور مثال کلمه “Java” ممکن است معنی

^۱ Hyperlinks

^۲ WEN

جزیره، کافه یا زبان برنامه نویسی باشد و از طرفی ممکن است کلمات کلیدی متفاوت مفاهیم یکسانی را در برگیرند مثل "بمب اتمی"، "پروژه منهن"، "سلاح هسته ای" و "ناگازاکی". لذا هر دو اصل یک و دو کنار یکدیگر استفاده می‌شوند.

از آنجایی که روش خوشه بندی مبتنی بر چگالی DBSCAN و نسخه‌ی افزایش‌اش، نیازمندی‌های این پژوهش را ارضا می‌کند، این روش‌ها در نظر گرفته شدند. در این پژوهش از داده‌های وب سایت (<http://encarta.msn.com/>) استفاده شد. Encarta شامل تعداد زیادی از اسنادی است که بر روی یک موضوع خاص به شکل سلسله مراتبی سازماندهی شده‌اند. مجموعه‌ی داده‌ها شامل نشست‌های جستار است که از لاگ‌های کاربر استخراج شده‌اند. در واقع هر نشست مطابق با یک جستار و اسنادی است که کاربر بر روی آن‌ها کلیک کرده است. قابل توجه است که در هر هفته حدوداً یک میلیون جستار صادر شده که حدود نیمی از نشست‌های جستار شامل کلیک روی اسناد هستند. خارج از این نشست‌ها حدود نود درصد دارای ۱ تا ۲ کلیک روی سند هستند. حتی اگر برخی از کلیک‌های سند اشتباه باشد، می‌توان انتظار داشت که بیشتر کاربران بر اسناد مرتبط کلیک کرده‌اند. ویرایشگران Encarta با مسئله‌ی تعیین سوالات پرتکرار کاربران روبرو بودند. هدف این پژوهش هم توسعه سیستم خوشه‌بندی برای سرعت بخشیدن به تعیین سوالات پرتکرار توسط ویرایشگران و ایجاد یک اصطلاحنامه آنلاین است که بتوان از طریق تحلیل لاگ کاربر، اصطلاحات استفاده شده توسط کاربران به اسنادی که این اصطلاحات در آنها یافت شدند را متصل نمود. لاگ‌های کاربر در بازه‌ی یکماه (حدود ۲۲ گیگابایت) از وب سایت Encarta جمع آوری شد. از این لاگ‌ها ۲۷۷۲۶۱۵ نشست جستار کاربر استخراج شد که از این بین ۲۰۰۰۰ نشست جستار جهت ارزیابی انتخاب شد نتایج نشان داد که ۴۹٪ از جستارها شامل تنها یک کلمه کلیدی و ۳۳٪ از جستارها شامل ۲ کلمه کلیدی هستند.

نتایج نشان داد که اغلب کاربران از یک مجموعه کوچک از کلمات استفاده کرده و اسناد را برای مطالعه انتخاب می‌کنند. همچنین این یافته، فرضیه پشت دیدگاه FAQ را تایید می‌کند که تعداد زیادی از کاربران به موضوعات مشابه (و یا پرسش‌های متداول FAQ) علاقمندند. تعداد زیادی از جستارهای مشابه در تعداد کمی از خوشه‌ها (بزرگترین خوشه‌ها) ظاهر می‌شوند. به عنوان مثال، ۴۵۰۰ جستار معروف (۲۲,۵٪ از کل تعداد جستارها) تنها به ۴۰۰ خوشه گروه‌بندی می‌شوند، که بیشتر مفهوم FAQ را تایید می‌کند. به این معنی که بسیاری از کاربران تمایل به استفاده از جستارهای مشابه در

طی یک دوره زمانی خاص دارند. در نهایت نتایج این پژوهش نشان داد که ترکیب کلمه‌ی کلیدی و لاگ‌های کاربر بهتر از استفاده از یکی از آن‌هاست.

در سال ۲۰۰۳، اوزمولتو^۱ و همکارانش [۳۹] نشان دادند که بسیاری از کاربران به خصوص در اینترنت جستجوهای کوتاهی انجام می‌دهند که نیازهای واقعی آن‌ها را به طور واضح نشان نمی‌دهد.

در سال ۲۰۰۴، بیتزل^۲ و همکارانش، صدها میلیون لاگ جستار یک سرویس جستجوی وب تجاری همه منظوره را از ۵۰ میلیون کاربر و در طول بازه‌ی زمانی یک هفته جمع‌آوری کردند [۱۰]. پژوهش آن‌ها بر روی بررسی جریان جستار-ها بر روی زمان تمرکز داشت. در این پژوهش تحلیل‌ها نشان داد که تغییرات در محبوبیت و یگانگی جستارهای به ساعتی که در طول روز جستجو می‌شوند، بستگی دارد. جهت بررسی ترافیک جستارها یک ساعت مشخص در نظر گرفته شد و ترافیک جستارها روی آن بررسی شد. نتایج پژوهش نشان داد که ترافیک جستارها از طبقه خاصی از موضوعات متفاوت با دیگر طبقه موضوعات و همینطور با کل جستارها متفاوت است. برای مثال، جستارهای مربوط به امور مالی افراد در ساعت ۷ تا ۱۰ صبح محبوب‌تر هستند. در حالی که جستارهای مربوط به موسیقی کمتر در این ساعات مورد تقاضا هستند. نتایج این پژوهش، دیدگاه ارزشمندی را جهت ارتقای اثربخشی بازیابی مسیریابی والگوریتم‌های ذخیره (caching) فراهم نمود.

رز و لوینسون در سال ۲۰۰۴، چارچوبی را برای فهم اهداف جستجوهای کاربران ارائه دادند [۴۳]. پژوهش‌های قبل از این کار، بیشتر بر روی فهم نحوه‌ی رفتار کاربران و اینکه آن‌ها چگونه و چه چیزی را جستجو می‌کنند تمرکز داشت نه بر روی اینکه چرا آن‌ها جستجو می‌کنند. جهت توسعه و آزمایش چهارچوب پیشنهادی، نمونه‌ای از جستارهای موتور جستجوی AltaVista مورد بررسی قرار گرفت. برای اینکه هدف اصلی هر کدام از جستارهای کاربر مشخص شود، باید امکان سوال از کاربر در مورد نیت و هدفش وجود داشته باشد، اما قطعاً این امکان در اکثر اوقات فراهم نیست. اما می‌توان هدف جستار را بطور ساده با نگاه کردن به خود جستار، نتایج برگردانده شده توسط موتور جستجو، نتایج کلیک شده توسط کاربر، جستجوهای بیشتر یا دیگر فعالیت‌های انجام شده توسط کاربر مشخص کرد. در واقع دو روش برای ارتباط

¹ Ozmutlu

² Beitzel

هدف‌ها با جستارها در زمان اجرا در موتورهای جستجو وجود دارد: کاربر می‌تواند به طور صریح هدف را از طریق واسط کاربری شناسایی کند، یا سیستم می‌تواند به صورت خودکار هدف را استنتاج نماید. ویژگی “I’m feeling lucky” که در موتور جستجوی گوگل هست، کمک می‌کند که کاربران به طور ضمنی هدف خود را به عنوان “حرکت به یک وب سایت خاص” اعلام کنند. این می‌تواند مثالی برای روش اول باشد. برای روش دوم یعنی طبقه‌بندی خودکار، می‌توان از روش‌های یادگیری ماشین یا روش‌های آماری استفاده کرد. در این روش‌ها صدها یا هزاران نمونه از مثال‌های جستارهای طبقه‌بندی شده (و ویژگی‌های مرتبط با آن‌ها) جهت آموزش نیاز است.

نتایج این پژوهش نشان داد که چارچوب پیشنهادی می‌تواند جهت ارتباط اهداف با جستارها استفاده شود. همچنین تحلیل نتایج، نشان داد که جستجوهای “هدایت شونده” کمتر از حد معمول رواج دارند در حالیکه “جستجوی منابع” ناشناخته بخش بزرگی از جستجوهای وب را شامل شود. از نتایج این پژوهش، یعنی درک اهداف کاربر از جستجو می‌توان جهت بهبود الگوریتم‌های موتورهای جستجو استفاده کرد. همچنین می‌توان شخصی‌سازی صفحات جستجو را متناسب با نیاز کاربر فراهم کرد. به طور مثال، نمایش تبلیغات مرتبط در زمینه خرید در زمانی که کاربر در حال جستجو برای خرید است، خوشایند است اما زمانی که مشغول جستجو برای انجام یک کار پژوهشی است، ناخوشایند است. در واقع “حساسیت به هدف” فاکتور مهمی در جستجوی کاربران است.

در سال ۲۰۰۵، یاتز^۱ و همکارانش [۹] بر روی تحلیل لاگ جستارهای مطرح شده در موتور جستجو تمرکز کردند. آن‌ها چندین مدل را استخراج کردند که بر روی چگونگی جستجوی کاربران و چگونگی استفاده از نتایج موتور جستجو تاکید داشتند. مدل‌های استخراج شده برگرفته از مدل مارکوف و با استفاده از تحلیل کلیک‌های کاربر و تحلیل عبارات مورد جستجو بدست آمده بود.

در پژوهش دیگری که در سال ۲۰۰۶ توسط اسپینک^۲ و جانسن [۴۶] انجام شد. آن‌ها دریافتند که کاربران دو یا سه واژه را جستجو کرده و همواره فقط اولین صفحه‌ی نتایج را مشاهده می‌کنند. در پژوهش دیگری که توسط آن‌ها

^۱ Yates

^۲ Spink

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۲۳

مفصل‌تر انجام گرفت [۲۷]، تعاملات بین کاربران و موتورهای جستجوی وب از دیدگاه طول نشست، طول جستار، پیچیدگی جستار و محتوای مشاهده شده در میان موتورهای جستجوی وب مقایسه شد. جهت مقایسه‌ی ۹ مطالعه‌ی بزرگ وب، برای ۴ موتور جستجوی وب اروپایی و ۵ موتور جستجوی وب ایالات متحده در طی یک دوره‌ی ۷ ساله بررسی شد. آن‌ها ۹ پژوهش را مورد بررسی قرار دادند:

(۱) موتور جستجوی وب (Spink, Jansen) Excite و (Saracevic, ۲۰۰۰)

(۲) موتور جستجوی وب (Hoßlscher و Strube, ۲۰۰۰)

(۳) موتور جستجوی وب (Silverstein) AltaVista و همکاران، (۱۹۹۹)

(۴) موتور جستجوی وب (Spink, Jansen, Wolfram) Excite و (Saracevic, ۲۰۰۱)

(۵) خدمات جستجوی وب (Cacheda و Vinqa, ۲۰۰۱، b۲۰۰۱)

(۶) موتور جستجوی وب (Spink) AlltheWeb.com و همکاران، (b۲۰۰۲)

(۷) موتور جستجوی وب (Spink, Jansen, Wolfram) Excite و (Saracevic, ۲۰۰۰)

(۸) AlltheWeb.com (اسپین و همکاران، b۲۰۰۲)

(۹) (Spink و Jansen) AltaVista (۲۰۰۲)

به طور خلاصه، این نه مطالعه ۲۸۷۲۱۲۰۰۰ (نزدیک به ۳۰۰ میلیون) نشست جستجوی وب و ۱۰۱۵۱۲۶۸۱۴ (بیش

از ۱ بیلیون) پرس‌وجو که مردم به موتورهای جستجوی وب ارسال می‌کنند، را ارائه می‌دهند.

شباهت‌ها و تفاوت‌هایی در میزان استفاده از موتورهای جستجوی بین افراد در ایالات متحده آمریکا و اروپا وجود

دارد. جستجوگران در هر دو، در طول نشست، طول جستار، و تعداد صفحات نتایج مشاهده شده مشابه هستند. استفاده از

اپراتورهای جستار وب در هر دو نسبتاً پایدار است اما میزان استفاده از اپراتورهای پیشرفته جستار در موتورهای جستجوی

ایالات متحده آمریکا بسیار بالاتر از اروپاست. نتایج نشان داد که برای موتورهای جستجوی وب ایالات متحده آمریکا،

درصد جستارهای تک واژه‌ای، در محدوده‌ی بین ۲۰ تا ۲۹ درصد از تمام جستارها است. برای کاربران موتور جستجوی

وب اروپایی، درصد جستارهای تک واژه‌ای عددی بین حدود ۲۵ تا ۳۵ درصد است.

تغییر در موضوعات جستجوی وب یکی دیگر از نتایج این تحقیق است. یافته‌های این پژوهش نشان می‌دهد که موضوعات درباره‌ی تجارت، مسافرت، اقتصاد، کامپیوتر، اینترنت و فناوری تقریباً ۲۵ درصد از پرس‌وجوها را تشکیل می‌دهند. درصد جستجوی وب در موضوعات جنسی در موتورهای جستجوی وب اروپایی و آمریکایی کاهش یافته است. روند کلی استفاده از وب به عنوان ابزاری برای کسب اطلاعات و تجارت و نه سرگرمی در حال افزایش است. این روند افزایشی، در جستجوگران ایالات متحده آمریکا در مقایسه با جستجوگران اروپایی بیشتر مشهود است. درصد جستجوگران مشاهده کننده‌ی فقط یک صفحه نتایج برای کاربران موتورهای جستجوی وب ایالات متحده و اروپایی در حال افزایش است.

در سال ۲۰۰۷، وایت^۱ و موریس^۲ لاگ تعامل کاربران حرفه‌ای و غیرحرفه‌ای موتور جستجو را مورد مطالعه قرار دادند [۵۰]. در واقع کاربران حرفه‌ای کاربران فعالی هستند که زمان کمتری را جهت جستجو و پیمایش دنباله جستجو صرف می‌کنند و انحراف کمتری در دنباله‌های جستجویشان وجود دارد و در جستجویشان موفق‌ترند. این پژوهش به بررسی تفاوت رفتارهای جستجو بین کاربران از لحاظ استفاده از دستور زبان پیشرفته تمرکز دارد. سوالات مطرح در این پژوهش این است که آیا رابطه‌ای بین استفاده از دستور زبان پیشرفته و دیگر مشخصه‌های جستجو وجود دارد؟ آیا رابطه‌ای بین استفاده از دستور زبان پیشرفته و رفتارهای هدایتی پس از اجرای جستارها وجود دارد؟ آیا رابطه‌ای بین استفاده از دستور زبان پیشرفته و معیارهای موفقیت جستجو وجود دارد؟

برای انجام پژوهش داده‌هایی از تعاملات لاگ کاربران در یک بازه ۱۳ هفته‌ای از ژانویه تا آوریل ۲۰۰۶ جمع‌آوری شد. اطلاعات شامل شناسه کاربر، زمان نمایش هر صفحه، یک مشخصه یکتا برای پنجره‌های مرورگر (برای حل این ابهام که برای نمایش یک صفحه از چه مرورگری استفاده شده)، و آدرس صفحه‌ی وب نمایش بازدید شده است. علاوه بر داده‌های جمع‌آوری شده در طول این مطالعه، یکسری داده که شامل قضاوت درباره ارتباط اسناد دیده شده توسط کاربران برای ۱۰۶۸۰ جستار منحصر به فرد موجود در لاگ‌های تعاملی بود، جمع‌آوری شد. این داوری‌ها در یک مقیاس شش نقطه‌ای

¹ White

² Morris

توسط داورهای آموزش دیده در زمان جمع‌آوری اطلاعات تعیین شد. در نهایت، لاگ‌های تعاملی ۵۸۶۰۲۹ کاربر منحصر به فرد را که میلیون‌ها جستار را بر روی سه موتور جستجوی محبوب گوگل، یاهو، و MSN اجرا کرده بودند، بررسی شد. به جز ویژگی‌های لاگ که البته دید کلی خوبی را از تعامل کاربران با موتورهای جستجو می‌دهند، دنباله‌های رفتاری کاربران نیز استخراج شد. برای انجام این کار، دنباله‌های جستجو از لاگ‌های تعاملی استخراج شد. یک دنباله جستجو^۱، یک سری از صفحات وب بازدید شده است که از طریق یک ابر لینک بهم متصل شده‌اند و با صفحه‌ی نتیجه جستجو آغاز شده و با یکی از وقایع زیر به پایان می‌رسند: هدایت به هر صفحه‌ای که از صفحه‌ی فعلی لینک داده نشده باشد، بستن صفحه‌ی مرورگر فعال، یا یک بازه‌ی نشست غیرفعال به مدت ۳۰ دقیقه. در حدود ۱۲٫۵ میلیون دنباله‌ی جستجو (حاوی حدود ۶۰ میلیون سند) از لاگ‌های تمامی کاربران استخراج شد. که میان‌ه‌ی تعداد دنباله‌های جستجو به ازای هر نفر ۳۰ و میان‌ه‌ی مراحل در دنباله‌ها ۳ بود. استخراج این دنباله‌ها امکان بررسی وضعیت رفتار جستجوی پس از اجرای جستارها را از طریق میانگین مدت زمان نشست جستجوی کاربران، میانگین مدت دنباله‌های جستجوی کاربران، میانگین زمان نمایش هر سند، میانگین تعداد گام در دنباله‌های جستجوی کاربران، تعداد شاخه‌ها در الگوهای مسیریابی کاربران و تعداد عملیات بازگشت در دنباله‌های جستجوی کاربران را فراهم نمود.

نتایج نشان داد که تفاوت‌های معنی داری در جستارها، کلیک بر روی نتایج، مرور کردن صفحات وب پس از اجرای جستار و جستجوی موفق کاربران حرفه‌ای نسبت به کاربران غیرحرفه‌ای وجود دارد. البته در مطالعات پیشین رابطه‌ی مستقیم بین دستور زبان جستارهای و میزان مهارت در جستجو به تنهایی مطالعه و تایید شده بود. در این پژوهش فرض بر این بود که کاربران تخصص لازم جستجو را دارا هستند و در این پژوهش هدف یافتن رابطه بین ویژگی‌های رفتاری و تخصص جستجو بود.

یافته‌های این پژوهش نشان می‌دهد که کاربران حرفه‌ای در حین جستجویشان چه نوع عملکردی داشته‌اند و چگونه از نوع تعاملات آن‌ها با سیستم می‌توان در جهت بهبود جستجوی تمامی کاربران با انواع سطوح استفاده کرد و استراتژی‌های جستجوی بهینه را به همه آموزش داد.

¹ Search trail

در سال ۲۰۰۸، گان^۱ و همکارانش پژوهشی را جهت بهبود جستجوهای جغرافیایی^۲ انجام دادند [۲۱]. آن‌ها علاقه-مند به کشف انواع جستارهای صادر شده توسط کاربران بودند. اینکه کاربران چه نوع جستارهای جغرافیایی (اطلاعاتی، هدایتی و تراکنشی) را صادر می‌کنند، آن‌ها در جستجوهایشان به دنبال چه نوع اطلاعاتی هستند و برای جستجوهای جغرافیایی‌شان چه نوع اصطلاحات جغرافیایی را به کار می‌برند. کاربران از چه سایت‌هایی به عنوان نتیجه‌ی یک جستار جغرافیایی بازدید می‌کنند. چگونه اصطلاحات جغرافیایی مختلف توسط کاربران مشابه به کار گرفته می‌شوند و اینکه چه اصطلاحات غیر جغرافیایی در رابطه با اصطلاحات جغرافیایی هستند. برای رسیدن با این اهداف آن‌ها یک طبقه‌بندی کننده، برای جداسازی جستارهای جغرافیایی و غیر جغرافیایی طراحی کردند. آزمایش خود را بر روی ۳۶ میلیون جستار که از روی موتور جستجوی AOL بدست آمده بود، انجام دادند. این داده‌ها از جستارهایی انجام شده توسط حدود ۶۵۰۰۰۰ کاربر در یک بازه زمانی سه ماه اول سال ۲۰۰۶ جمع آوری شد. داده‌ها شامل شناسه کاربر (به صورت ناشناس)، جستار، زمانی که جستار صادر می‌شود، رتبه نتایج کلیک شده^۳ و نتیجه که کاربر کلیک کرده^۴ می‌باشند. جستارها در ۴ طبقه برچسب‌گذاری شدند:

(۱) جستارهای جغرافیایی که حاوی نام شهرستان، ایالت یا کشور به عنوان یک اصطلاح جغرافیایی هستند

(۲) جستارهای جغرافیایی که حاوی چنین اصطلاحاتی نیستند

(۳) جستارهای غیر جغرافیایی که به نظر می‌رسد حاوی عبارتی جغرافیایی هستند مانند "مرغ سوخاری کنتاکی"، "نیویورک تایمز" و یا "اولین بانک نیاگارا"

(۴) جستارهای غیر جغرافیایی بدون اصطلاحات جغرافیایی

نتایج نشان می‌دهد که طبقه‌بندی کننده طراحی شده با دقت حدود ۹۶٫۶۹٪ همه‌ی جستارهای را به درستی طبقه-

بندی می‌کند. همچنین یافته‌های پژوهش نشان می‌دهد که تفاوت‌های مهمی بین جستارهای جغرافیایی و غیر جغرافیایی وجود دارد. جستارهای جغرافیایی اغلب هدفشان در مکان یابی و دریافت خدمات می‌باشد. در حالیکه جستارهای

¹ Gan

² Geographic

³ Item-Rank

⁴ ClickURL

غیرجغرافیایی بیشتر هدفشان سرگرمی، دانلود یا مشاهده صفحات اطلاعاتی است. همچنین جستارهای جغرافیایی به استفاده از تعداد واژه‌های بیشتری نسبت به جستارهای غیرجغرافیایی تمایل دارند.

در این پژوهش طبقه‌بندی‌کننده دیگری طراحی شد که متفاوت با قبلی عمل می‌کند. یعنی به جای اینکه واژه‌های جستار را مد نظر قرار دهد، کلیک کاربران روی نتایج را مد نظر قرار می‌دهد. این طبقه‌بندی‌کننده با دقت ۸۷٫۹۴٪ جستارها را طبقه‌بندی نمود. نتایج نشان می‌دهد که کاربران به صورت تکراری جستجوهای را روی نواحی جغرافیایی مشابه انجام می‌دهند. در واقع به سادگی می‌توان شهر کاربران را با توجه به اصطلاحات جغرافیایی که جستجو می‌کنند، استخراج کرد. زیرا نتایج نشان می‌دهد که که کاربران اکثرا تمایل به جستجو برای سرویس‌های محلی دارند.

در سال ۲۰۰۸، دوپرت^۱ و پیوواسکی^۲، تلاش کردند که با در نظر گرفتن یک سری مفروضات در مورد رفتار کاربر در حین مرور صفحات وب، احتمال دیده شدن یک سند را تخمین بزنند [۱۸]. جهت آموزش، آزمایش و مقایسه این مدل از مجموعه بزرگی از داده‌های واقعی که شامل مجموعه‌ای از جستارها از یک موتور جستجوی تجاری در بازه زمانی چندین ماهه بود، استفاده شد. از آنجایی که نیاز به داده‌های معتبر جهت مدل‌سازی می‌باشد. لذا هر کدام از جستارهایی که کمتر از ۱۰ نشست داشتند و همه زوج‌های جستار-سندی را که کمتر از ۱۰ بار مشاهده شده بودند، کنار گذاشته شدند. همچنین جستارهایی که کمتر از ۵ کلیک در هر نشست داشتند، حذف شدند چون این جستارها احتمال خطای زیادی دارند و یا خیلی مبهم هستند. نهایتاً ۵۴۲۶۵۱ جستار که در ۳۶۴۳۶۸۰۸ نشست رخ داده بودند، انتخاب شد. قابل ذکر است احتمال اینکه یک سند کلیک شود، بستگی به مرتبط بودن آن به جستار و همچنین عواملی مانند موقعیت آن در صفحه جستجو دارد. در لیست ۱۰ جستجوی اول، احتمال کلیک، به ترتیب از بالا به پایین کم می‌شود. همچنین آزمایشات ردگیری چشم^۳ نشان داده است که کاربران به مشاهده نتایج انتهای صفحه تمایلی ندارند. نتایج این پژوهش نشان داد که روش پیشنهادی این پژوهش، به میزان معنی‌داری نسبت به مدل‌های قبلی بهتر عمل کرده است. یافته‌های پژوهش تایید کننده این است که یک کاربر تقریباً همیشه سندی را که دقیقاً بعد از سندی که کلیک کرده بود قرار دارد، را مرور می‌کند.

¹ Dupret

² Piwowarski

³ Eye tracking

در سال ۲۰۰۹، سیلوستری^۱ در پژوهشی که انجام داد بر روی اصول، الگوریتم‌های پایه‌ای و تکنیک‌هایی که برای استخراج دانش مفید از منابع نامحدود اطلاعاتی یعنی جستارهای کاربران تمرکز کرد [۴۵]. در این پژوهش، تکنیک‌هایی که با لاگ‌های جستار سروکار دارند جهت افزایش عملکرد موتورهای جستجوی وب مورد بررسی قرار گرفت. تکنیک‌هایی از جمله یادداشت‌های تاریخی و مقدماتی^۲، بسط جستار، پیشنهاد جستار، شخصی‌سازی نتایج جستار، یادگیری رتبه‌بندی و تصحیح املای جستار مورد بررسی قرار گرفت.

در سال ۲۰۰۹، بندرسکی^۳ و کرافت^۴ بر روی تحلیل جستارهای بلند (طولانی) در لاگ‌های جستجوی کاربران یک موتور جستجوی تجاری با هدف شناسایی ویژگی‌های پرتکرار در این نوع جستارها تمرکز کردند [۱۲]. آن‌ها به دنبال درکی از نوع نگارش جستارهای بلند توسط کاربران، بررسی رفتار آن‌ها پس از دریافت نتایج و تلاش جهت بهبود بازیابی نتایج جستارهای بلند بودند.

جهت تجزیه و تحلیل جستارها سه ویژگی طول، نوع جستارها و فرکانس آن‌ها در نظر گرفته شد: طول جستار همان طول رشته جستجو و یا در واقع تعداد توکن‌های کلمه (دنباله‌ای از کاراکترها با فاصله) است. نوع جستار به دو دسته بلند و کوتاه تقسیم می‌شوند. اکثر جستارها در لاگ جستجو کوتاه هستند. جستارهایی با طول کمتر از ۴ حدود ۹۰٫۳ درصد کل جستارها را تشکیل می‌دهند و برای ۹۹٪ جستارها، طول جستارها کمتر از ۱۲ است. جستارهای بلند طولشان بین ۵ تا ۱۲ است. جستارهای بلند به دسته‌های سوالی، اپراتوری، ترکیبی و غیرترکیبی (عبارت اسمی و عبارت فعلی) تقسیم می‌شوند. همچنین در این پژوهش، داده‌های کلیک در لاگ‌های جستجو، با هدف درک بهتر رفتار کاربر و عملکرد بازیابی موتور جستجو برای جستارهای طولانی مورد بررسی قرار گرفت.

نتایج نشان می‌دهد که اولاً کاربران در زمان مشاهده‌ی لیست نتایج جستار تمایل دارند نتایج پایین لیست را برای جستارهای طولانی‌تر نسبت به جستارهای کوتاه‌تر کلیک کنند. ثانیاً در رفتار کلیک کردن کاربران بر روی نتایج بدست آمده از انواع مختلف جستارهای بلند، تفاوت وجود دارد.

¹ Silvestri

² Historical Notes and Preliminaries

³ Bendersky

⁴ Croft

براساس بررسی‌های پژوهش‌های پیش از این پژوهش توسط داون^۱ و همکارانش، واضح است که جستارهای ناقص (کمیاب و کمتر درخواست شده) کمتر کلیک و بازدید می‌شوند و جستارهای دوباره بازنویسی شده پرتکرار هستند [۱۷]. داون و همکارانش معتقدند که فرکانس جستار بسیار مهمتر از طول جستار است، اما در نتایج این پژوهش مشاهده شد که فرکانس و طول جستار با یکدیگر برای بازیابی نتایج بهینه موثر هستند.

در سال ۲۰۱۰، تیلر^۲ و تیوان^۳ پژوهشی را جهت بازیابی مجدد^۴ نتایج موتورهای جستجو انجام دادند [۴۷]. اینکه چگونه جستارهای بازیابی مجدد متفاوت با جستارهای سنتی^۵ عمل می‌کنند، مورد توجه این پژوهش قرار گرفت. آن‌ها بر روی سه منبع داده با مقیاس بزرگ تمرکز کردند:

(۱) لاگ‌های جستار (جستارها، کلیک‌ها، نتایج)

(۲) لاگ‌های مرورگرهای وب (مشاهده URL)

(۳) جستجوی روزانه وب (محتوای صفحه)

داده‌های جمع‌آوری شده جهت انجام این پژوهش از ماه ژانویه ۲۰۰۹ در سه بخش جمع‌آوری شد. در بخش لاگ‌های جستار موتور جستجو، تقریباً ۹۰۰ میلیون کلیک نتایج جستجو از ۱۶۰ میلیون کاربر بر روی موتور جستجوی Live Search (now Bing) انجام شد. در بخش لاگ‌های مرورگر وب، اطلاعات ردیابی مسیرهای دنبال شده توسط ۴ میلیون کاربر و شامل صدها میلیون صفحه بازدید شده بود؛ جمع‌آوری شد و برای بخش محتوای صفحه، ۵۵۰۰۰ صفحه متفاوت نمونه‌برداری شدند.

ویژگی‌های بازیابی مجدد مطالعه شده بر روی داده‌ها شامل دو بخش در موتور جستجو و صفحات هستند: در موتور جستجو، طول جستار، محبوبیت، نتایج رتبه‌بندی در کلیک، و در بخش صفحات، محتوای صفحات و لینک دنبال شده اولیه.

^۱ Downe

^۲ Tyler

^۳ Teevan

^۴ Re-finding

^۵ Traditional

تفاوت‌ها بین جستارهای قبلی و جستارهای مجدد بازیابی شده و آن‌هایی که تغییرات کوچکی داشتند، بررسی شد. نتایج نشان داد که جستارهای مجدد بازیابی شده، وقتی تغییرات اساسی هستند، نسبت به جستارهای قبلی که استفاده شده‌اند، تمایل بیشتری به بهتر شدن دارند. جستارهای بازیابی مجدد، به طور معمول کوتاه‌تر هستند و URL‌هایی که مجدد یافت می‌شوند، رتبه بالاتری دارند. همچنین یافته‌های این پژوهش تاکید دارند که زمانی که کاربری روی نتیجه قبلی کلیک می‌کند، احتمال دارد که مسیر مشابهی را مانند سایر کاربرانی که بر همان URL کلیک کرده‌اند دنبال کند. همچنین جستارهای بازیابی مجدد تمایل به همگرایی دارند. وقتی کاربری به صورت مداوم از یک موتور جستجو برای پیدا کردن نتایج یکسان استفاده می‌کند، جستار استفاده شده ممکن است در مراحل اولیه متفاوت باشد اما در نهایت سازگار شده و به ثبات می‌رسد. تمامی یافته‌های این پژوهش جهت بهبود عملکرد موتورهای جستجو می‌توانند مورد استفاده قرار گیرند.

در سال ۲۰۱۲، سندیل^۱ و همکارانش یک شمای^۲ طبقه‌بندی‌کننده برای رفتار تایپ کردن کاربران پیشنهاد دادند[۱۴]. هدف آن‌ها بررسی چگونگی تحلیل لاگ جستارهای یک سیستم جستجوی فوری و یافتن رفتارهای کاربر در جستجوی فوری است. از آنجا که هدف سیستم‌های اطلاعاتی این است که به کاربران اجازه دهند که با سرعت بالا نتایج جستجوهایشان را بیابند. جستجوی کلمه‌ی کلیدی روش پذیرفته شده‌ای برای رسیدن به این هدف است و برای جستجوی مجموعه‌ای از اسناد استفاده می‌شود. در روش سنتی مبتنی بر جستجوی کلمه کلیدی، کاربر یک جستار کامل را به سیستم صادر می‌کند و نتایج جستجو مرتبط با آن برگشت داده می‌شوند. این الگوی جستجو نیازمند این است که کاربر بتواند یک جستار کامل و معقول را ارسال کند تا نتایج مرتبط با آن را بیابد. زمانی که کاربر دانش محدودی درباره‌ی مسئله‌ای دارد وقتی جستار را ارسال می‌کند، مجبور است با عملیات رفت و برگشتی و آزمون و خطا در دریافت نتایج، جستار را اصلاح کند تا نتیجه مطلوب را اخذ کند. اکثر موتورهای جستجو و تعدادی از فرم‌های جستجوی آنلاین از روشی به نام جستجوی فوری یا لحظه‌ای^۳ پشتیبانی می‌کنند. جستجوی فوری روی کاراکتر به کاراکتر جستار که کاربر تایپ کرده، جستار را به عنوان نتیجه پیشنهاد می‌دهد. در سیستم‌های پیشنهاد فوری جستارهای مربوط به جستار کاربر به طور مداوم پیشنهاد می‌شوند.

¹ Cetindil

² Scheme

³ Instant

برای مثال، اگر کاربر عبارت "hom" را تایپ کند، سیستم چندین جستار مشابه مانند homologous و home care را پیش‌بینی و پیشنهاد می‌کند و سپس کاربر می‌تواند یکی از این پیشنهادات را انتخاب و آن جستار را به موتور جستجو ارسال کند و نتایج را بازیابی کند.

سیستم‌های جستجوی فوری می‌توانند به دو طبقه تقسیم‌بندی شوند: سیستم‌های پیشنهاد فوری^۱ و سیستم‌های نتایج فوری^۲. سیستم‌های پیشنهاد فوری بر لاگ‌های جستار برای استخراج جستارهای معروف (پراستفاده) تکیه دارند. سیستم‌های نتایج فوری، نتایج جستجو را متناسب با انواع کاربر پیشنهاد می‌کنند. برای مثال، واسط کاربری در Netflix^۲ به کاربر امکان جستجوی فیلم‌ها را براساس ویژگی‌های عنوان، بازیگران، کارگردان و ژانرها می‌دهد. اگر یک کاربر جستار جزئی را تایپ کند، سیستم فیلمی را که با آن کلمه کلیدی به عنوان پیشنهاد منطبق می‌شود، نمایش می‌دهد. به طور مثال بعد از اینکه "bat" تایپ شد، "Batman Forever" و "Battlestar Galactica" نشان داده می‌شوند.

روش پیشنهادی این پژوهش بر روی لاگ‌های جستار سیستم جستجوی PSearch، انجام شد. تعداد ۳۵۰۰۰۰ جستار که توسط ۲۸۰۰ کاربر در یک دوره یکساله، از سپتامبر ۲۰۱۰ تا سپتامبر ۲۰۱۱ ارسال شده بود مورد بررسی قرار گرفت. رکوردهای لاگ، در سه مرحله تقسیم‌بندی شدند: براساس آدرس مکان دسترسی کاربر^۳ آن‌ها به جستارهای جداگانه از کاربرهای مختلف تقسیم شدند. سپس این مجموعه‌ها با توجه به تفاوت‌های زمانی بین جستارهای متوالی گروه-بندی شدند. اگر تفاوت زمانی بیشتر از ۵ دقیقه بود آن جستار، به نشست دیگری نسبت داده شد. در نهایت، گروه‌های مرحله قبلی براساس شباهت فاصله-ویرایش^۴ جستارها به منظور قراردادن جستارهای متوالی مشابه در یک نشست و جداسازی جستارهایی که توسط افراد مختلف صادر شدند، تقسیم‌بندی شدند.

^۱ Instant-suggestion system

^۲ Instant-result system

^۳ IP

^۴ edit-distance

در نهایت پژوهشگران توانستند لاگ جستارهای یک سیستم جستجوی فوری را تحلیل و نشست‌ها را براساس رفتارهای تایپ کردن کاربران طبقه‌بندی کنند. مزایای جستجوی فوری با جستجوی سنتی براساس عواملی چون میزان تلاش کاربر، زمان مورد نیاز برای برطرف کردن نیاز اطلاعاتی و نرخ موفقیت مقایسه شد. نتایج نشان داد که جستجوی فوری می‌تواند به طور معمول ۲ ثانیه زمان را در هر جستجو صرفه‌جویی کند و همچنین تلاش کاربر را برای تایپ کردن را کاهش و همچنین باعث افزایش نرخ موفقیت در بازگرداندن نتایج مورد نظر کاربر شود.

در سال ۲۰۱۲، پژوهش دیگری [۵۳] جهت تحلیل ۷۵۰ میلیون جستار بدست آمده از یک موتور جستجوی تجاری واقعی انجام شد. طول جستار، نرخ پالایش جستار، نرخ کلیک، نرخ جستار از طریق جستجوی پیشرفته از فاکتورهای مورد بررسی در این پژوهش بودند. نتایج نشان داد که برخی از رفتارهای کاربران همیشه ثابت است. به عنوان مثال بیش از ۹۰٪ کاربران تنها به ۵ نتیجه اولی که موتور جستجو برمی‌گرداند توجه دارند. همچنین نتایج نشان داد با توجه به موضوع مورد جستجو، کاربران رفتارهای متفاوتی را نسبت به نتایج بازیابی جستجو نشان می‌دهند.

در سال ۲۰۱۳، توجه به کاربران و ویژگی‌های منحصربه‌فرد هر کاربر افزایش یافت. به طوری‌که در پژوهشی که انجام گرفت [۱]، نتایج نشان داد که بین سبک‌شناختی و رفتار جستجوی اطلاعات کاربران رابطه وجود دارد و افراد با سبک شناختی کلامی از تعداد واژه‌های بیشتری نسبت به افراد با سبک شناختی تصویری استفاده کرده‌اند.

در سال ۲۰۱۴، پژوهشگران دانشگاه برکلی پژوهشی را بر روی بیش از ۲۰۰۰۰۰ جستارهای بدست آمده از Splunk انجام دادند [۷]. درواقع Splunk یک بستر نرم‌افزاری^۱ قدرتمند جهت جمع‌آوری لاگ‌ها، جستجو، مشاهده، آنالیز و تحلیل داده‌هاست. این بستر با پردازش هزاران لاگ تولید شده توسط نرم‌افزارها، مفسرها و تجهیزات امنیتی، اطلاعات ارزشمندی را که با چشم قابل رویت نیستند، آشکار می‌کند. در این پژوهش هدف پژوهشگران بررسی رفتار تحلیل لاگ به صورت کمی بود تا بتوانند از نتایج آن در تولید ابزارهای تحلیل لاگ استفاده نمایند.

در واقع یکی از چالش‌هایی که سیستم‌های لاگ با آن مواجه‌اند این است که نمایشی از فعالیت‌های انسانی تحلیل لاگ وجود ندارد. برای کمک به این مسئله، در این پژوهش یک پرسشنامه طراحی شد و در اختیار کسانی که از Splunk

¹ Platform

استفاده می‌کنند قرار داده شد. سؤالاتی از قبیل شما معمولاً چه نوع تبدیلاتی را بر روی داده‌های لاگ جهت تحلیل آن‌ها انجام می‌دهید؟ روند معمول تحلیل به چه شکل است؟ چه کسی کار تحلیل لاگ را انجام می‌دهد و در آخر به چه نتیجه‌ای می‌رسید؟ چه پیشرفت‌هایی در ابزارهای تحلیلی نیاز است تا آن‌ها عملکرد خوبی داشته باشند و فهم فرآیند تحلیل برای کاربر هم بهتر شود و به سادگی بتوانند داده‌هایشان را استخراج کنند. نتایج بررسی پاسخ کاربران به این سؤالات در سه بخش فیلترینگ، فرمت مجدد و خلاصه‌سازی داده‌ها بررسی شد. براساس مشاهدات پژوهشگران این پژوهش، حجم زیادی از تحلیل لاگ در Splunk مربوط به فیلترینگ است. دلیلش این است که تحلیل لاگ اغلب برای مسائلی استفاده می‌شود که به دنبال رویدادهای غیرنرمال و یا بخش خاصی از داده‌ها هستند، به طورمثال برای عیب‌یابی حساب کاربری، اشکال زدایی عملکرد، تشخیص نفوذ و سایر مشکلات مربوط به امنیت. یک دلیل دیگر این است که اطلاعات بسیار زیادی که در لاگ‌ها جمع آوری شده‌اند، اکثراً در زمان توسعه‌ی سیستم و جهت اشکال زدایی سیستم ذخیره شده‌اند و برای کاربر نهایی مفید و کارا نیستند. از طرفی از آنجایی که بیشتر داده‌های لاگ شده به صورت ناسازگاری سازماندهی شده‌اند. بخش منظور فرمت مجدد و پاکسازی داده‌ها نیز بخش بزرگی از فعالیت تحلیل لاگ است. خلاصه‌سازی نیز یک تکنیک بسیار پرکاربرد در تحلیل داده‌هاست و برای ایجاد انواع نمودارها و گراف‌ها جهت نمایش به کاربر مفید باشد. در تحلیل لاگ با Splunk خلاصه‌سازی با در نظر گرفتن بعد زمان یکی از مهمترین موارد استفاده است.

۳-۲- تحلیل لاگ موتورهای جستجوی پایگاه‌های علمی

پژوهش‌های ذکر شده در بالا، اطلاعات زیادی در مورد رفتار جستجو در وب در اختیار پژوهشگران می‌گذارد. اما بی‌شک، رفتار جستجو در موتورهای جستجو با رفتار جستجو در یک پایگاه تخصصی علمی متفاوت است. از این‌رو برخی از پژوهش‌ها تنها به بررسی رفتارهای جستجو در پایگاه‌های علمی پرداختند.

در سال ۲۰۰۰، یو^۱ و اپس^۲ از داده‌های لاگ برای بررسی رفتار کاربر در پروژه SuperJournal استفاده کردند [۵۶]. آن‌ها داده‌های لاگ بازه‌ی زمانی ۲۳ ماه (فوریه ۱۹۹۷ تا دسامبر ۱۹۹۸) را که شامل ۱۰۲،۹۶۶ فعالیت ثبت شده کاربران بود، بررسی کردند. فعالیت‌ها در ارتباط با چهار خوشه‌ی موضوعی، ۴۹ مجله، ۸۳۸ شماره‌ی مختلف یک مجله، ۷۸۶،۱۵ مقاله و سه موتور جستجوی وب در ارتباط بودند. متغیرهایی مانند تعداد نشست‌ها برای هر مجله، تعداد نشست‌ها در هر ماه، فرکانس استفاده و زمان استفاده بررسی شدند. نتایج این پژوهش نشان می‌دهد که لاگ‌ها می‌توانند منابع مقرون به صرفه و خوبی برای استخراج اطلاعات از نحوه‌ی تعامل کاربران با سیستم باشند اما آن‌ها نمی‌توانند دلیل رفتار کاربران را شرح دهند، لذا منابع اطلاعاتی دیگر برای بررسی رفتار کاربران مورد نیاز است.

در سال ۲۰۰۲، کیا^۳ و همکارانش [۵۷] لاگ فایل رفتاری کاربران ScienceDirect در Elsevier را برای بازه‌ی ۱ ژانویه تا ۱۸ سپتامبر ۲۰۰۰ بررسی کردند که اطلاعات کتابشناختی و متن کامل مقالات بیش از ۱۳۰۰ مجله با حدود ۶۲۵۰۰۰ کاربر را در بردارد. گزارش‌های آماری از جمله تعداد بازدیدها در یک دوره زمانی، موضوعات مورد جستجو، تعداد جستار برای هر مکان دسترسی کاربر یافته‌های این پژوهش است.

در سال ۲۰۰۳، وانگ و همکارانش، ۵۴۱۹۰۲۰ جستار کاربر بر روی یک موتور جستجوی وب سایت دانشگاهی را در یک دوره ۴ ساله (می ۱۹۹۷ تا می ۲۰۰۱) بررسی کردند [۴۸]. تمرکز آن‌ها بر روی شناسایی مشکلاتی از جمله تایپ غلط جستارها توسط کاربران و پیشنهاداتی جهت رفع آن‌ها، فهم رفتاری جستارهای کاربر با توجه به فرکانس‌های تکرار جستجو و توسعه تکنیک‌هایی جهت تحلیل جستارها بود. یافته‌های این پژوهش نشان داد که الگوهای مشابهی در طول سال-ها در بین رفتار جستجوی کاربران وجود دارد. جستارهای انجام شده توسط کاربران، براساس موضوع و تعداد جستجو، دارای تغییرات فصلی هستند. از دیگر یافته‌های این پژوهش، تحلیل‌های آماری است که نشان می‌دهد ۳۸ درصد از تمام

¹ Yu

² Apps

³ Kea

جستارها تنها شامل یک واژه هستند و اکثر جستارها منحصر به فرد می‌باشند. همچنین از دیگر یافته‌های این پژوهش مشاهده‌ی مدل‌های ذهنی کاربران در حین جستجو و یادگیری از سیستم از طریق اصلاح درخواست‌ها است.

در سال ۲۰۰۶، یی^۱ و همکارانش [۵۵] در بررسی رفتار جستجوی کاربران، بر روی داده‌های پایگاه روانشناسی PsycINFO و دو پایگاه حوزه‌ی تاریخ با نام ABC-Clio و America تمرکز کردند. ۱۹۶۸۴۴۸ رکورد از پایگاه داده PsycINFO و ۱۰۳۱۰۰۰ رکورد از دو پایگاه داده دیگر مورد بررسی قرار گرفتند. پژوهشگران این پژوهش دریافتند که ترکیبی از واژه‌ها، شناخت بهتری از رفتار نگارش جستار کاربران را در اختیار قرار می‌دهد. همچنین نتایج این پژوهش نشان داد که میانگین طول جستار در پایگاه داده‌های روانشناسی PsycINFO، ۳،۱۶ واژه و در دو پایگاه دیگر ۳،۴۲ واژه بوده است. نتایج نشان داد علیرغم اینکه طول جستارها کوتاه می‌باشد، اصطلاحات چند واژه‌ای که شامل دو یا سه واژه هستند اغلب در جستارها استفاده می‌شوند و نرخ رخداد آن‌ها در PsycINFO ۰،۷۲ و در ABC-Clio ۰،۹۳ است.

یافته‌ی دیگر این پژوهش این است که واژه‌های انتخاب شده در جستارها به شدت به موضوع مورد جستجو وابسته است. برای مثال در تاریخ، جستجوگران بیشتر به استفاده از تعداد زیادی اصطلاحات چند واژه‌ای غیر مفهومی مثل افراد، نواحی و رویدادها تمایل دارند درحالی‌که در روانشناسی، کاربران تمایل به استفاده از اصطلاحات چند واژه‌ای مفهومی مثل درمان، اختلال و رشد دارند.

در سال ۲۰۰۸، نیکلاس و همکارانش [۳۶] به بررسی لاگ پایگاه علمی ScienceDirect پرداختند. هدف آن‌ها انجام تحلیل عمیقی از لاگ کاربران و فهم دقیقی از رفتار جستجوی اطلاعات آنان بود. برای دستیابی به این هدف، پرسشنامه‌ای توسط کاربران سیستم تکمیل شد که کاربران را با توجه به موضوع مورد جستجو آن‌ها مشخص می‌کرد. لاگ-های رفتار جستجوی کاربران در یک دوره ۱۸ ماهه جمع‌آوری شد که شامل ۱۶۸۶۵ نشست و ۱۱۰۰۲۹ صفحه‌ای بود که توسط کاربران بازدید شده بودند. داده‌های بیش از ۷۵۰ کاربر معتبر شناخته شد. نتایج تحلیل بر روی این داده‌ها نشان داد که میان رفتار جستجوی افراد در رشته‌های مختلف علمی، تفاوت‌های زیادی وجود دارد که این رفتار شامل تعداد دفعات بازدید و تعداد جستجو در یک نشست بود. برای مثال محققان علوم اجتماعی، بیشتر از سایرین تنها چکیده مقاله یا مقالات

¹ Yi

در حال انتشار را مشاهده می‌کردند. در حالیکه محققان علم مواد در یک نشست به طور متوسط اکثر مجلات منحصر به فرد را مشاهده کردند. یافته‌های این پژوهش می‌تواند جهت بهبود پایگاه‌های علمی مانند ScienceDirect به منظور ارائه سرویس مناسب‌تر برای کاربران در موضوعات مختلف مورد استفاده قرار گیرند.

در سال ۲۰۱۱، ژیاندونگ^۱ و همکارانش [۲۹] ویژگی‌هایی از جمله توزیع زمانی دسترسی کاربران، توزیع طول عبارات جستجو و نسبت بازیابی اطلاعات پایگاه داده‌ی علمی-تخصصی مجلات چینی را تحلیل کردند. آن‌ها همچنین استراتژی جستجوی کاربران دانشگاه‌های مختلف را مورد بررسی قرار دادند. نتایج نشان داد که کاربران با رشته‌های مختلف و سطوح علمی متفاوت، رفتارهای متفاوتی را در جستجو از خود بروز می‌دهند. در نهایت، براساس نتایج این پژوهش، یک مدل استراتژی جستجوی آکادمیک کاربران دانشگاهی بدست آمد.

در سال ۲۰۱۳، پارک^۲ و لی^۳ [۴۰] بر روی لاگ یک وبگاه علمی تحلیلی را انجام دادند که هدفشان کسب دانشی در خصوص رفتار کاربر جهت بهبود سیستم‌های بازیابی اطلاعات بود. آن‌ها داده‌های ۷۲۴۰۰۰۰ جستار را برای دوره‌ی یکساله، سال ۲۰۱۰ بررسی کردند. یافته‌های این پژوهش نشان می‌دهد که اغلب جستارها در علم و تکنولوژی کوتاه و ساده هستند که نشان از این است که کاربران تمایل دارند حداقل تلاش را در جهت جستارها انجام دهند. یافته‌های دیگر این پژوهش نشان می‌دهد اغلب نشست‌های کاربران کمتر از ۳۰ دقیقه طول کشیده، میانگین آن‌ها ۲۴ دقیقه و ۱۵ ثانیه، و حداقل آن‌ها ۱ ثانیه و حداکثر ۱۰ ساعت بوده است.

در سال ۲۰۱۴، ادانتن^۴ [۳۸] در پایان نامه‌ی کارشناسی ارشد خود، لاگ فایل سامانه‌ی بازیابی اطلاعات کتابخانه دانشگاه Strathclyde را در بازه‌ی زمانی دو ماهه مورد بررسی قرار داد. وی در این پژوهش، جستارهای کاربران و الگوهای جستجو را مورد تحلیل قرارداد. نتایج بدست آمده از این پژوهش در جهت بهبود طراحی واسط کاربر و سازماندهی محتوا به کار گرفته شد.

¹ Jiandong

² Park

³ Lee

⁴ Oduntan

ساتو و همکارانش اخیراً بر روی استفاده از لاگ‌ها و تحلیل آن جهت یافتن رابطه بین کاربران، مسیرهای دسترسی و مقالات دستیابی شده توسط آن‌ها تمرکز کردند [۶۰]. آن‌ها از پنج پایگاه داده ژاپنی جهت جمع‌آوری لاگ‌ها استفاده کردند و داده‌های جمع‌آوری شده برای یک دوره یکساله از ژانویه ۲۰۰۸ تا دسامبر همان سال است. نتایج این پژوهش نشان می‌دهد که تنها پژوهشگران نیستند که از منابع علمی استفاده می‌کنند بلکه تعداد افراد غیر محقق بسیار بیشتر است و بازه‌ی کاربرانی که در موتورهای جستجو به دنبال اطلاعات علمی هستند، بسیار وسیع است. همچنین از دیگر یافته‌های این پژوهش این است از آنجایی که دسترسی به منابع علوم کاربردی از علوم پایه بیشتر است، لذا یادگیری افراد در این حوزه‌ها بیشتر است.

پارک و لی در سال ۲۰۱۶، پژوهشی را بر روی ۲۴۸۲۰۴۱۶ جستار که توسط کاربران در سال ۲۰۰۸ تا ۲۰۱۱ ارسال شده بود، انجام دادند [۵۸]. هدف آن‌ها فهم طولی^۱ تعاملات کاربر و سیستم در حوزه‌ی علم و تکنولوژی بود. نتایج نشان داد که رفتار جستجوی کاربران و الگوهای رفتاری آنان ثابت است و آن‌ها تمایل به استفاده از جستارهای ساده و کوتاه دارند. همچنین عباراتی که در سال‌های ۲۰۰۸ تا ۲۰۱۱ زیاد مورد جستجو بودند، استخراج شدند. زمان تصمیم‌گیری برای کاربران پس از صدور جستار به طور میانگین ۱۰ ثانیه گزارش شد. به طور متوسط زمان ۱۰ ثانیه ثبت شده است که آیا کاربران می‌خواهند روی نتایج بدست آمده از جستار تمرکز کنند یا خیر. اگر کاربران قرار باشد نتایج بدست آمده را مشاهده کنند به طور میانگین زمان حدود ۱۰ دقیقه را برای این کار صرف می‌کنند.

همان‌طور که پیشینه‌ی پژوهش نشان می‌دهد با وجود اینکه تحلیل لاگ به عنوان یکی از شیوه‌های ارزیابی سامانه‌های اطلاعاتی و همچنین مطالعه‌ی رفتار اطلاع‌یابی کاربران در خارج از ایران جایگاه خود را پیدا کرده است. اما در ایران کار پژوهشی چندانی در این زمینه انجام نشده است. تعدادی پژوهش نظری در زمینه تحلیل لاگ انجام شده است [۲] [۶] [۴] [۵] و در این میان، تنها دو پژوهش عملی بر روی تحلیل لاگ‌های پایگاه اطلاعات و مدارک علمی ایران (ایرانداک) و رفتار جستجوی کاربران آن در انجام گرفته است [۳] [۵۹]. در پژوهشی که در سال ۱۳۹۲ انجام گرفت یک سری تحلیل‌های آماری بر روی یک دوره ۲۵ روزه لاگ جستجوی کاربرانی که به جستجوی اطلاعات در این پایگاه

¹ longitudinal

اطلاعاتی پرداخته‌اند، انجام شد [۳]. نتایج این پژوهش روزهای پر استفاده و پربسامدترین طول عبارت جستجو و موضوع جستجو را مشخص می‌کند. در این پژوهش تنها به ارائه یکسری گزارش‌های آماری اکتفا شده است که می‌تواند شروع‌کننده خوبی در مسیر این دست پژوهش‌ها باشد اما تحلیلی بر روی رفتارهای کاربران صورت نگرفته است. در پژوهشی که اخیراً انجام گرفت [۵۹]، محققان لاگ جستجوی کاربران را برای یک بازه طولانی‌تر (فوریه ۲۰۱۳ تا جولای ۲۰۱۵) را که شامل ۱۸۹۴۳۸ رکورد می‌باشد، بررسی نمودند. تمرکز آن‌ها تنها بر روی جستجوهای انجام شده در علوم محیطی بود. در این پژوهش از تکنیک متن کاوی و پردازش زبان طبیعی استفاده شد. هدف محققان بررسی وضعیت پژوهش‌های مورد تقاضای پژوهشگران علوم محیطی در ایران و یافتن خلاهای پژوهشی در این حوزه بود. در نهایت، نتایج نشان داد که حوزه‌هایی مانند آلودگی، مراعات و مدیریت از حوزه‌هایی هستند که در پژوهش‌های ایرانی دارای خلا هستند. از طرفی پژوهش‌های کافی در حوزه‌های جنگل، خاک و صنایع در ایران انجام شده است.

۴-۲- نتیجه‌گیری

در این بخش، پژوهش‌هایی که تاکنون در حوزه‌ی تحلیل لاگ انجام شده است، مورد بررسی قرار گرفت. این پژوهش‌ها در دو دسته‌بندی ارائه شد. یکی تحلیل لاگ موتورهای جستجوی عمومی و دیگری تحلیل لاگ موتورهای جستجوی پایگاه‌های علمی. در بخش، تحلیل لاگ موتورهای جستجوی پایگاه‌های علمی، پژوهش‌هایی که بر روی پایگاه‌های علمی خارجی و داخلی انجام شده است، بررسی شد.

فصل سوم

روش انجام پژوهش

۱-۳- مقدمه

سامانه‌ی ملی اطلاعات پایان نامه‌ها/رساله‌های دانش‌آموختگان داخل کشور تحت عنوان سامانه‌ی گنج در پژوهشگاه علوم و فناوری اطلاعات ایران، پیاده‌سازی شده است. بخش پایگاه‌داده‌ی آن با استفاده از نرم‌افزار MySQL پیاده‌سازی شده است. کاربران از طریق این سامانه قادر به جستجوی اطلاعات مورد نیاز خود هستند. نمایی از سامانه در شکل ۱ آمده است.



شکل ۱- نمایی از سامانه گنج

۲-۳- روش پژوهش

در زمانی که کاربران از موتور جستجوی این سامانه استفاده می‌کنند، اطلاعات جستجوی آنان نیز توسط نرم‌افزار MySQL ثبت و ذخیره می‌شود. نمونه‌ای از لاگ‌های ذخیره شده در شکل ۲ آمده است:

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۴۱

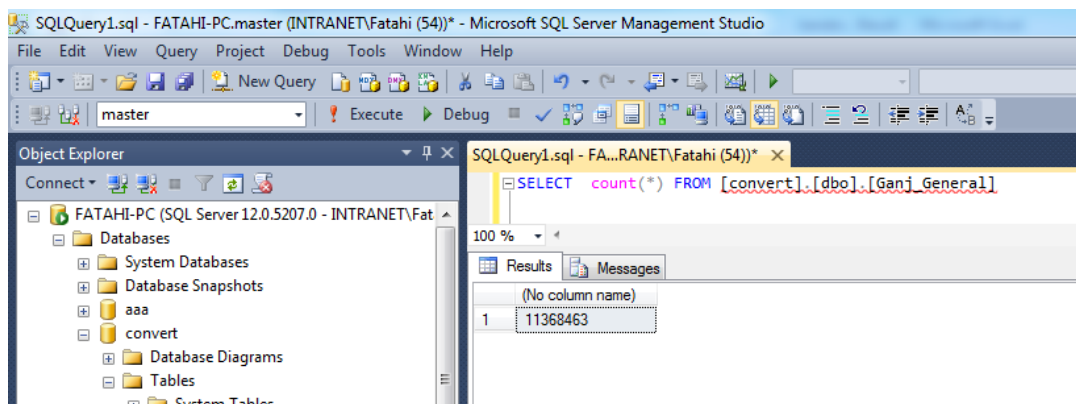
	query	user_id	created_at	number	ip
1	الگوی مصرف	375827	2016-09-22 00:00:03	107	66.249.73.129
2	فن آوری اطلاعات	375827	2016-09-22 00:00:03	37	66.249.73.129
3	صور بیانی	375827	2016-09-22 00:00:04	99	5.160.134.165
4	برنامه توسعه	375827	2016-09-22 00:00:06	16552	66.249.73.129
5	معیار آزمون	375827	2016-09-22 00:00:08	3115	157.55.39.81
6	حمل و نقل دریایی	375827	2016-09-22 00:00:08	280	207.46.13.49
7	ساختار جنگل	375827	2016-09-22 00:00:10	11	157.55.39.7
8	استخراج با حلال	375827	2016-09-22 00:00:12	47	157.55.39.7
9	همی پلازی	375827	2016-09-22 00:00:16	55	66.249.73.129
10	(آهن+۲)	375827	2016-09-22 00:00:17	2583	157.55.39.197
11	سخت فسیلی	375827	2016-09-22 00:00:18	1870	66.249.73.129
12	تاب آوری	375827	2016-09-22 00:00:20	465	66.249.73.129
13	دست رفتن فرصت	375827	2016-09-22 00:00:23	77	188.34.160.35
14	...الترانه دانش و عوامل فنا	375827	2016-09-22 00:00:25	41	66.249.73.129
15	مسیریابی	375827	2016-09-22 00:00:26	206	66.249.73.129
16	ارزش افزوده اقتصادی	375827	2016-09-22 00:00:28	49	207.46.13.49
17	کارگران کارخانه	375827	2016-09-22 00:00:29	14	207.46.13.49
18	مدیریت پرسناری	375827	2016-09-22 00:00:33	250	66.249.73.129
19	دست رفتن فرصت	375827	2016-09-22 00:00:35	77	188.34.160.35

شکل ۲- نمایی از لاگ موتور جستجوی گنج

قابل ذکر است، فایل لاگ موتور جستجو سامانه گنج شامل اطلاعاتی از قبیل شناسه کاربر، عبارت جستجو، تعداد دفعات جستجو، مکان دسترسی کاربر، زمان و تاریخ جستجو است. در این پژوهش بررسی‌ها بر روی لاگ‌های سامانه‌ی گنج در سه ماه سوم سال ۱۳۹۵ انجام گرفت. حجم رکوردهای این لاگ فایل ۱۱۲۵۲۹۷۴ تعداد است. با استفاده از زبان SQL، جستارهایی بر روی این لاگ فایل اجرا شد و اطلاعات مورد نیاز استخراج شد. در واقع، یک فایل خام لاگ در اختیار این پژوهش قرار گرفت، که با استفاده از زبان SQL، جستارهایی جهت استخراج نگارش شد. نمونه‌ای از آن‌ها در ادامه به همراه خروجی آمده است.

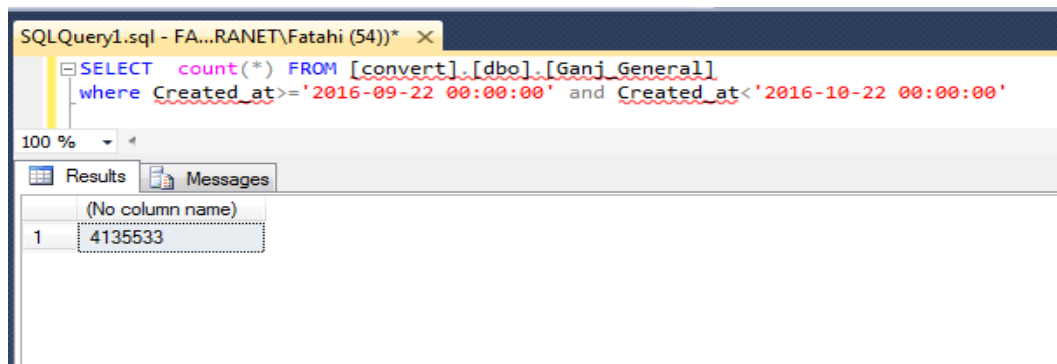
یک پایگاه داده به نام Ganj_General داریم که کل داده‌ها در آن موجود است. جستار زیر جهت شمارش تعداد کل رکوردهای پایگاه داده Ganj_General نگارش شده است.

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۴۲



شکل ۳- جستار جهت شمارش تعداد کل رکوردها

جستار شکل ۴، جهت شمارش رکوردها در بازه‌های جداگانه هر ماه نگارش شده است:



شکل ۴- جستار جهت شمارش تعداد رکوردها در بازه‌های جداگانه هر ماه

جهت شمارش فرکانس عبارات جستجو شده در سامانه گنج و همچنین بازه‌های جستجو، جستار شکل ۵ و ۶ نگارش شد.

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۴۳

SQLQuery1.sql - FA...RANET\Fatahi (54))* X

```
SELECT DISTINCT query, COUNT(query) FROM [convert].[dbo].[Ganj_General]
GROUP BY query
ORDER BY COUNT(query) DESC;
```

100 %

Results Messages

	query	(No column name)
1	الض سازی اینوگووین	15616
2	دانشگاه علوم پزشکی خدمات بهداشتی درمانی تهران	8302
3	دانشگاه تهران	7619
4	دانشکده ادبیات و علوم انسانی	6024
5	معماری	5974
6	حسابداری	5621
7	دانشگاه تربیت مدرس	5074
8	دانشکده علوم پایه	4595
9	دانشگاه علوم پزشکی و خدمات بهداشتی درمانی تهران	3864
10	دانشکده علوم تربیتی و روانشناسی	3787
11	مدیریت	3748
12	فرهنگ بومی	3568
13	دانشکده ادبیات علوم انسانی	3528
14	داده کاوی	3495
15	بازاریابی	3355
16	دانشکده علوم	3298
17	علوم انسانی	3268

شکل ۵- جستار جهت شمارش فرکانس عبارات جستجو شده در سامانه گنج

```

-- Create table [convert].[dbo].New_Saat12_14
([Query] nvarchar(255)
, [Number] int);

-- insert into [convert].[dbo].New_Saat12_14
SELECT query, DATEPART(HH, created_at) FROM [convert].[dbo].[new1]
where DATEPART(HH, created_at) >= '12' and DATEPART(HH, created_at) < '14'

```

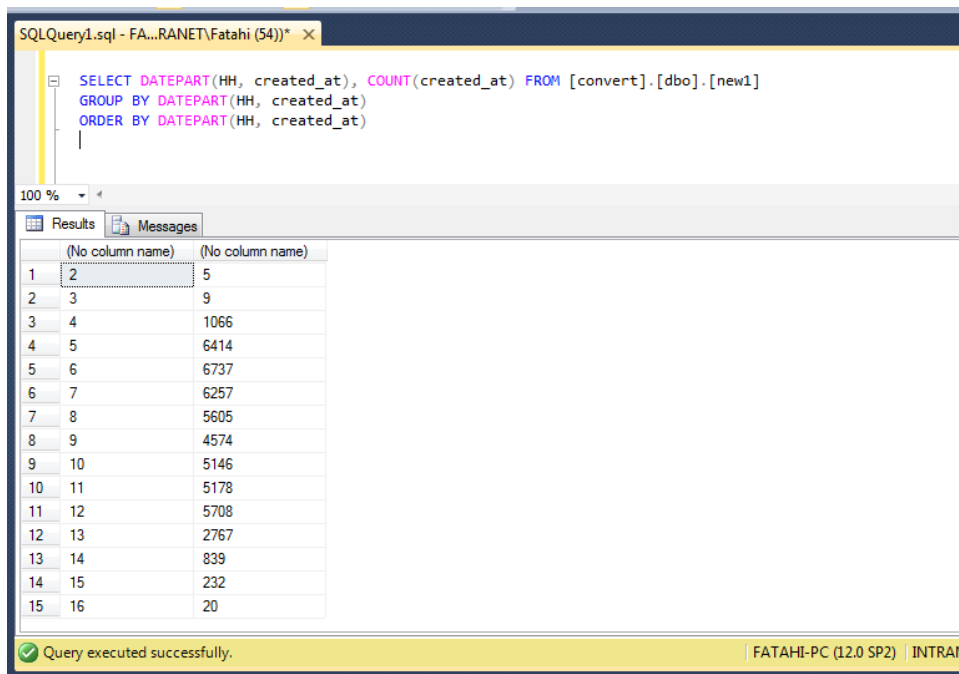
100 %

Results Messages

	(No column name)
1	28
2	30
3	33
4	28
5	25
6	24
7	24
8	26
9	26
10	26
11	18
12	18
13	24
14	21

شکل ۶- جستار جهت شمارش فرکانس عبارات جستجو شده در ساعت ۱۲ تا ۲ بعدازظهر در سامانه گنج

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۴۴



SQLQuery1.sql - FA...RANET\Fatahi (54))

```
SELECT DATEPART(HH, created_at), COUNT(created_at) FROM [convert].[dbo].[new1]
GROUP BY DATEPART(HH, created_at)
ORDER BY DATEPART(HH, created_at)
```

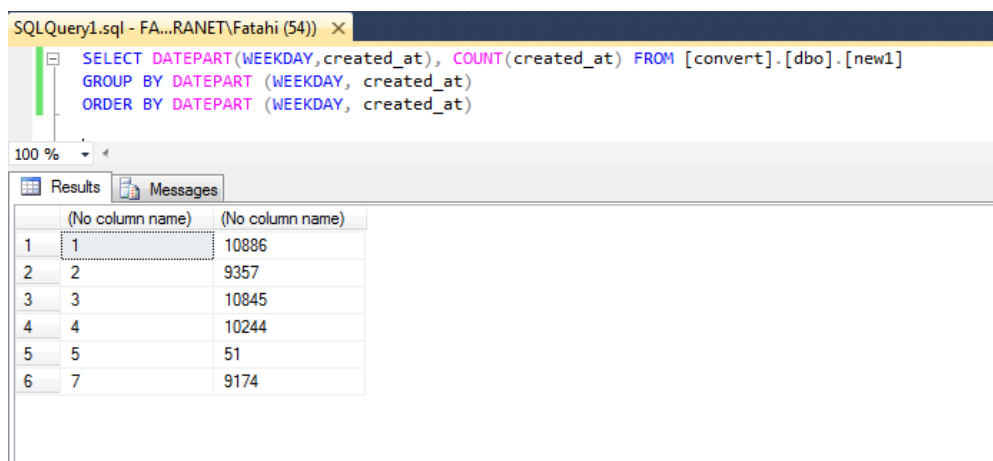
100 %

Results Messages

	(No column name)	(No column name)
1	2	5
2	3	9
3	4	1066
4	5	6414
5	6	6737
6	7	6257
7	8	5605
8	9	4574
9	10	5146
10	11	5178
11	12	5708
12	13	2767
13	14	839
14	15	232
15	16	20

Query executed successfully. FATAHI-PC (12.0 SP2) INTRAM

شکل ۷- جستار جهت استخراج تعداد جستجوها براساس روزهای هفته



SQLQuery1.sql - FA...RANET\Fatahi (54))

```
SELECT DATEPART(WEEKDAY, created_at), COUNT(created_at) FROM [convert].[dbo].[new1]
GROUP BY DATEPART(WEEKDAY, created_at)
ORDER BY DATEPART(WEEKDAY, created_at)
```

100 %

Results Messages

	(No column name)	(No column name)
1	1	10886
2	2	9357
3	3	10845
4	4	10244
5	5	51
6	7	9174

شکل ۸- جستار جهت استخراج تعداد جستجوها براساس ساعات

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۴۵

query	(No column name)
1 معماری	123
2 مدیریت دانش	91
3 آموزش مجازی	63
4 ایران	40
5 علم اطلاعات و دانش‌شناسی کتابداری و اطلاع رسانی	36
6 اقتصاد مقاومتی	34
7 برق	30
8 نقشه دانش	26
9 آموزش الکترونیکی	25
10 فیزیک	25
11 فلسفه	23
12 (مثنوی) کتاب	22
13 اسناد	19
14 امکان‌سنجی آموزش مجازی	19
15 فرهنگ & رسانه	19
16 جنگ ایران عراق	18

شکل ۹- جستار جهت استخراج کلیدواژه‌های پرتکرار

با استفاده از جستارهای نگارش شده در sql server نسخه‌ی ۲۰۱۴، اطلاعات قابل استخراج از داده‌های خام

لاگ فایل بدست آمد و با استفاده از نرم افزار Excel نسخه‌ی ۲۰۱۰ نمودارهای مربوط رسم شده است که در بخش نتایج به آن‌ها پرداخته شده است.

قابل ذکر است که لاگ فایل مورد بررسی در این پژوهش، در قالب سه ماه مهر، آبان و آذر تفکیک شد. جهت

تحلیل بهتر و منطقی‌تر لاگ فایل جستجوی کاربران را در سه حالت بررسی کرده‌ایم. حالت اول کل جستجوهای که بر

روی سامانه‌ی گنج انجام گرفته است و حالت دوم تنها جستجوهای مورد بررسی قرار گرفت که از طریق تالار جستجوی

پژوهشگاه علوم و فناوری اطلاعات ایران صورت گرفته است، یعنی جستجوی کاربرانی که به صورت حضوری و از داخل

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۴۶

پژوهشگاه اقدام به جستجو کرده‌اند. لذا تمام جستارهای نگارش شده بر روی هر دو نسخه‌ی داده‌ها اجرا شد و نتایج بدست آمد. حالت سوم، بررسی جستجوهای داخل کشور مورد نظر قرار گرفت.

۳-۳- نتیجه‌گیری

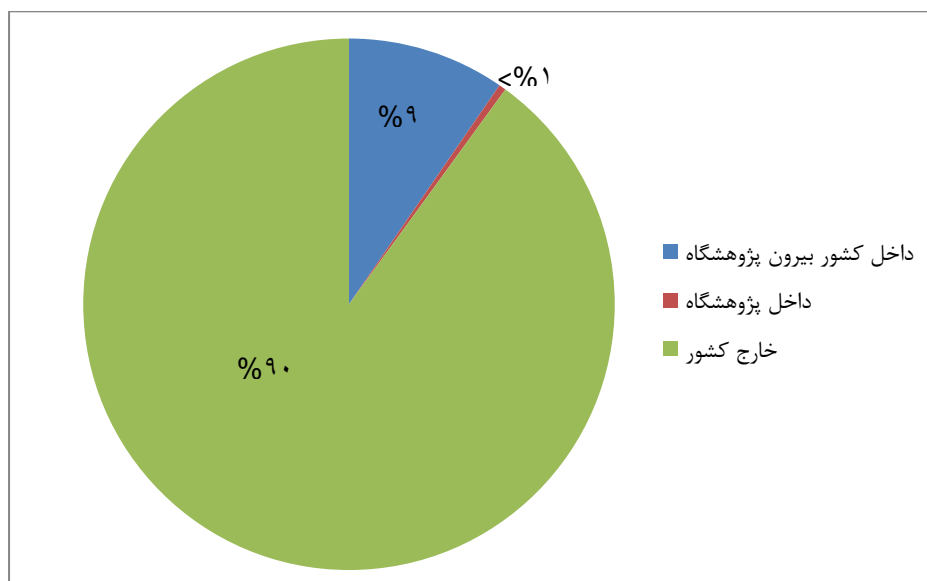
در این بخش، محدود مشخصی از داده‌های لاگ فایل انتخاب شدند، جستارهای مورد نیاز جهت استخراج داده‌ها نگارش شده و داده‌های مورد نیاز جهت تحلیل لاگ آماده‌سازی شدند.

فصل چهارم

نتایج و تحلیل آنها

۱-۴- مقدمه

لاگ فایل مورد بررسی در این پژوهش، در قالب سه ماه مهر، آبان و آذر تفکیک شد. جهت تحلیل بهتر و منطقی‌تر، لاگ فایل جستجوی کاربران را در سه حالت بررسی کرده‌ایم. حالت اول کل جستجوهای که بر روی سامانه‌ی گنج انجام گرفته است و حالت دوم تنها جستجوهای مورد بررسی قرار گرفت که از طریق تالار جستجوی پژوهشگاه علوم و فناوری اطلاعات ایران، یعنی جستجوی کاربرانی که به صورت حضوری و از داخل پژوهشگاه اقدام به جستجو کرده‌اند. حالت سوم بررسی جستجوهای است که از داخل کشور صورت گرفته‌اند. قابل ذکر است که ۹۰ درصد جستجوها با ip خارج از کشور و حدود ۱۰ درصد از جستجوها به داخل کشور مربوط است که از این میزان رقمی کمتر از یک درصد به ip های داخل پژوهشگاه مربوط است و نشان می‌دهد که جستجوها به صورت حضوری انجام گرفته است (شکل ۱۰).



شکل ۱۰- نسبت جستجوهای داخل، خارج از کشور و داخل پژوهشگاه

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۴۹

با توجه به بازه ip های موجود که از داخل پژوهشگاه جستجو انجام می‌شود. میزان جستجو توسط ip های داخل پژوهشگاه در جدول ۱ آمده است.

جدول ۱- میزان جستجوها توسط ip های داخل پژوهشگاه

جستجو	ip های داخل پژوهشگاه
۵۰۳۱۳	10.19.157.202
۱۸۹	10.19.157.204
۴۳	194.225.91.24
۵	46.209.199.24
۲	46.209.199.10
۲	10.19.157.205
۱	194.225.91.123
۱	194.225.91.233
۱	46.209.199.58

همچنین با توجه به رنج ip های موجود که از داخل کشور جستجو انجام می‌شود. میزان جستجو توسط ip های داخل کشور در جدول ۲ آمده است.

جدول ۲- میزان جستجو توسط ip های داخل کشور

سازمان	ISP	تعداد دفعات جستجو	IP
دانشگاه	Information Technology Company (ITC)	۸۴۳۸	217.218.171.16
دانشگاه	Information Technology Company (ITC)	۷۹۱۳	78.38.150.10
دانشگاه فردوسی مشهد	Mihan Communication Systems Co.,ltd	۷۸۹۱	46.143.248.48
شرکت فناوری اطلاعات (ITC)	Information Technology Company (ITC)	۷۱۲۷	195.146.37.19
دانشگاه شهید مدنی آذربایجان	Information Technology Company (ITC)	۵۸۲۵	85.185.1.110
دانشگاه فردوسی مشهد	Mihan Communication Systems Co.,ltd	۵۲۶۸	46.143.248.50
دانشگاه گیلان	Information Technology Company (ITC)	۵۲۴۹	78.38.47.10
شرکت فناوری اطلاعات (ITC)	Information Technology Company (ITC)	۵۲۱۸	217.218.215.40
دانشگاه محقق اردبیلی	Information Technology Company (ITC)	۴۷۱۵	80.191.193.34
دانشگاه تبریز	Information Technology Company (ITC)	۴۶۲۴	80.191.200.170

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۵۰

نتایج جدول ۲ نشان می‌دهد در مورد برخی از ip ها تعیین مکان دقیق امکان پذیر نبوده است. اما به طور کلی اکثر جستجوهای داخل کشور از دانشگاه‌ها انجام شده است و این امر کاملاً بدیهی و طبیعی است. همچنین دانشگاه‌های فردوسی مشهد، شهید مدنی آذربایجان، گیلان، محقق اردبیلی و تبریز در صدر جستجوگران سامانه‌ی گنج هستند و این نتایج به نتایج بدست آمده از طرح خانم خسروی نزدیک است. نکته قابل توجه در اینجا این است که روند استفاده از سامانه‌ی گنج در طی چندین سال بین برخی از دانشگاه‌ها یکسان است و تنها برخی از دانشگاه‌ها مایل به استفاده از این سامانه هستند. تغییر سامانه‌ی گنج براساس نیاز دانشگاه‌های دیگر می‌تواند به استفاده بیشتر این سامانه توسط دانشگاه‌های بزرگ کمک نماید.

نتایج بدست آمده بر روی تحلیل داده‌های لاگ فایل سه ماهه‌ی سوم سال ۱۳۹۵ سامانه گنج به شرح زیر است:

جدول ۳- آمار تعداد جستجوها به تفکیک ماه

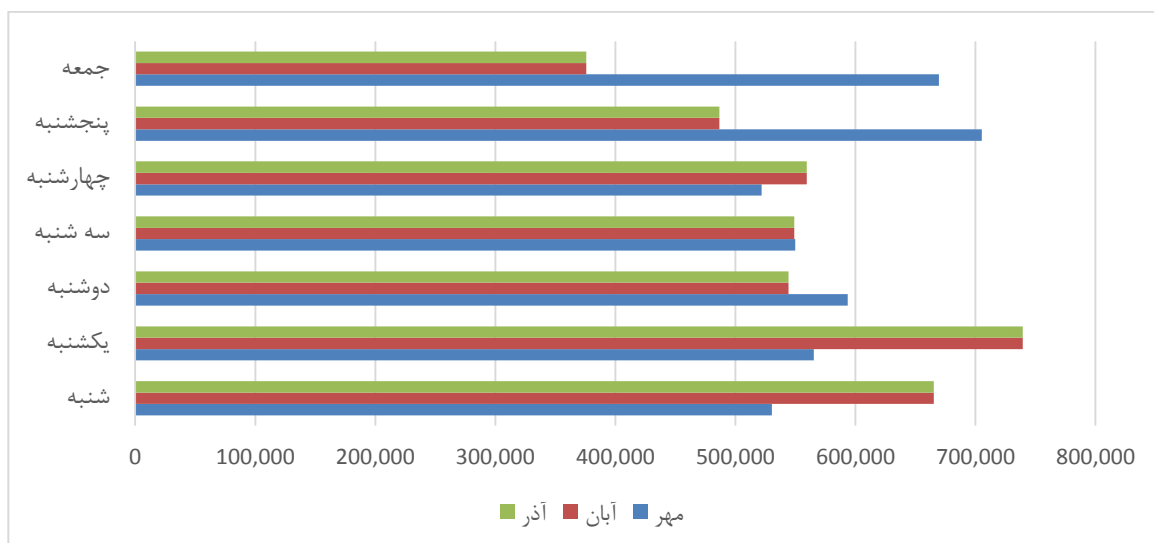
بازه جستجو	تعداد جستجوها با ip های داخل پژوهشگاه	تعداد جستجوها با ip های داخل کشور	تعداد جستجوهای کل
مهر ماه ۱۳۹۵	۵,۳۱۶	۳۰۵,۱۴۳	۴,۱۳۵,۵۳۳
آبان ماه ۱۳۹۵	۱۷,۱۹۲	۴۴۰,۲۰۴	۳,۹۱۹,۹۰۱
آذر ماه ۱۳۹۵	۲۶,۳۸۱	۳۸۱,۷۶۶	۳,۱۹۷,۵۴۰
تعداد کل	۴۸,۸۸۹	۱,۱۲۷,۱۱۳	۱۱,۲۵۲,۹۷۴

همانطور که مقادیر جدول ۳ نشان می‌دهد بیشترین آمار جستجو کلی مربوط به مهرماه ۱۳۹۵ می‌باشد. اما این تعداد برای جستجوهای حضوری در آذرماه به بیشینه مقدار خود رسیده است که دلیل آن را می‌توان نزدیکی به انتهای ترم تحصیلی، ارائه پروپزال و دفاع از رساله در نظر گرفت.

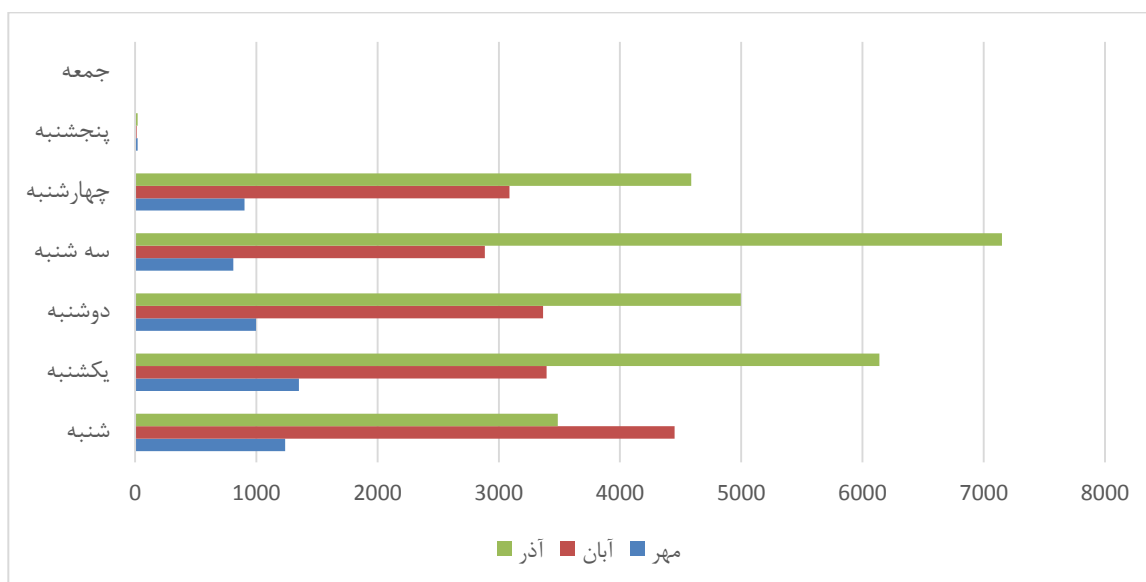
۲-۴- تحلیل داده‌ها براساس زمان و تاریخ

با نگاهی و اجرای یک جستار بر روی فایل لاگ در نرم افزار Sql server حجم جستجوهای انجام شده بر روی سامانه‌ی گنج برای هریک از روزهای هفته در هر یک از ماه‌های مهر، آبان و آذر به تفکیک بدست آمد که در شکل ۱۱ و ۱۲ به طور کامل نمایش داده شده است.

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۵۱



شکل ۱۱- میزان کل جستجوها به تفکیک روزهای هفته



شکل ۱۲- میزان جستجوها از داخل پژوهشگاه به تفکیک روزهای هفته

روزهای پنجشنبه در مهرماه بالاترین آمار جستجوی کاربران را گزارش می‌دهد و این حالت در روزهای یکشنبه

در آبان و آذر ماه مشاهده می‌شود و روزهای انتهای هفته در آبان و آذرماه از نرخ جستجوی بسیار کمی نسبت به سایر

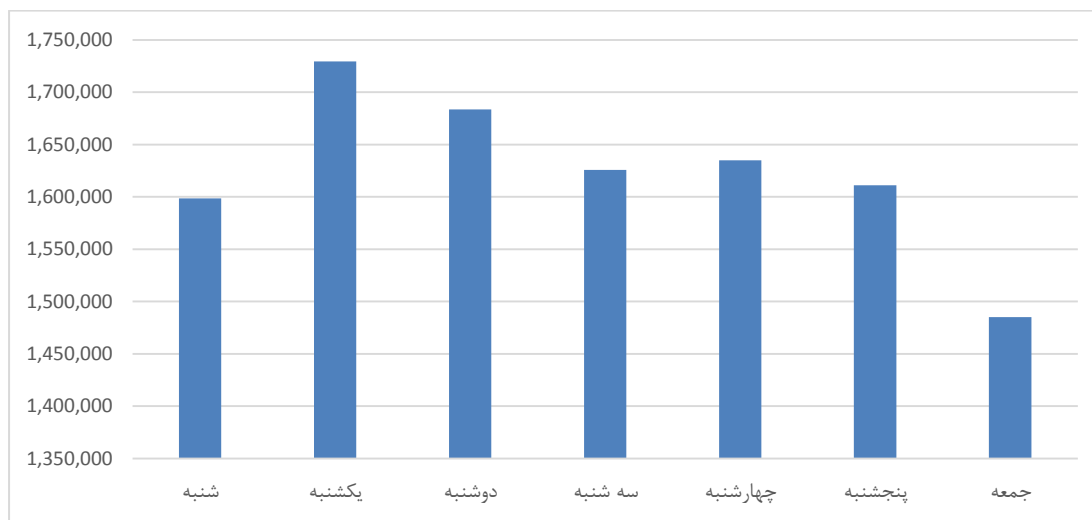
تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۵۲

روزها برخوردار هستند (شکل ۷). در حالی که در جستجوهای حضوری با توجه به تعطیلی پژوهشگاه در روزهای پنجشنبه و

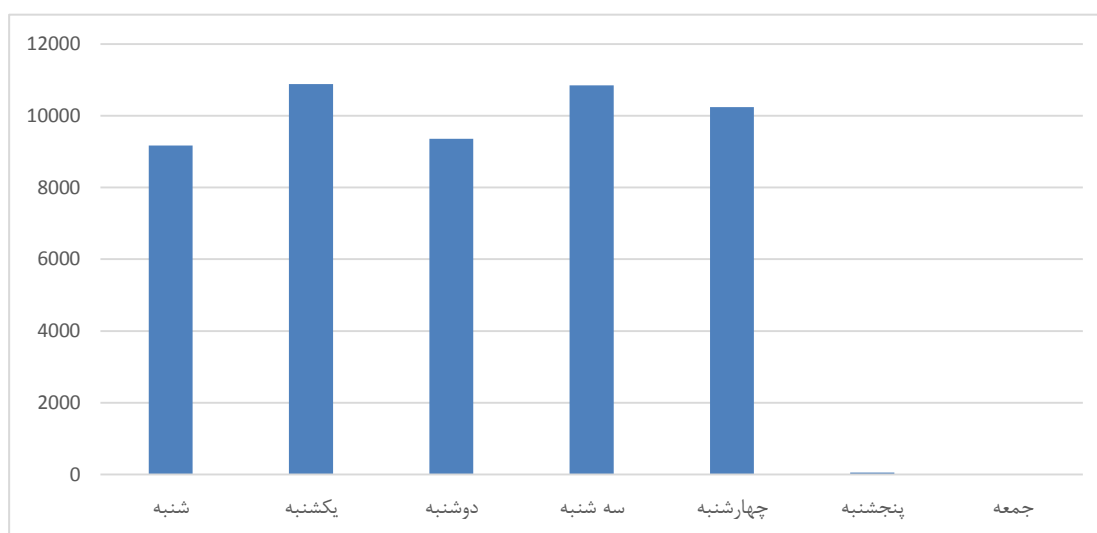
جمعه، بالاترین حجم جستجو مربوط به سه شنبه در آذرماه و روزهای شنبه در آبان ماه و روز یکشنبه در مهرماه است.

حجم جستجوی انجام شده بر روی سامانه‌ی گنج به طور کلی براساس روزهای هفته در شکل ۱۳ و ۱۴ آمده

است:

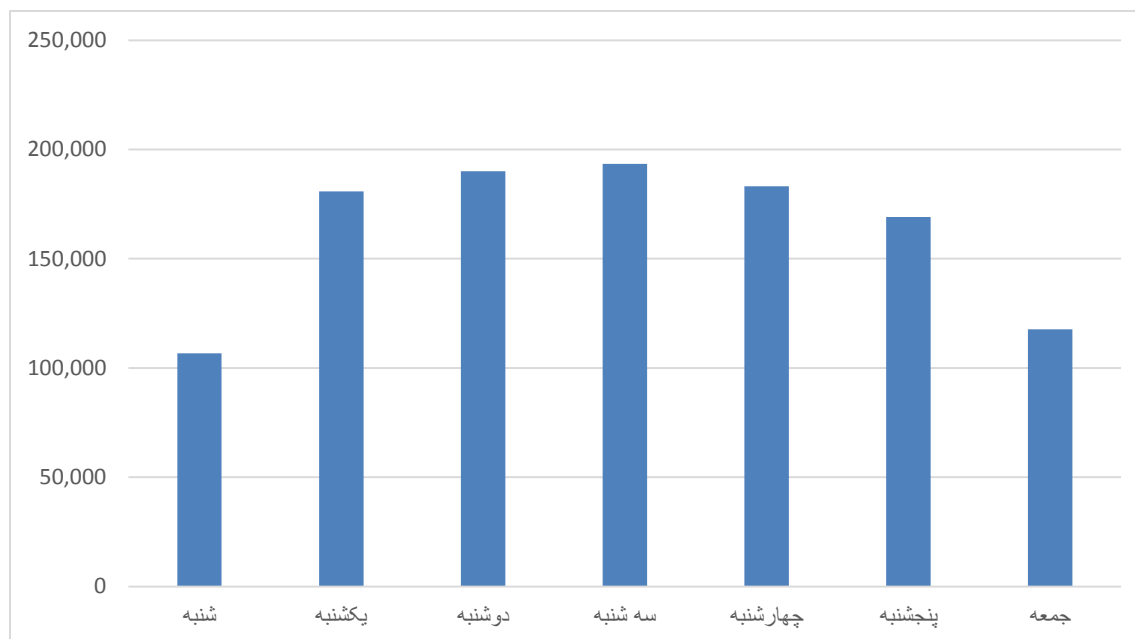


شکل ۱۳- حجم جستجوی انجام شده بر روی سامانه‌ی گنج به طور کلی براساس روزهای هفته



شکل ۱۴- حجم جستجوی انجام شده بر روی سامانه‌ی گنج از داخل پژوهشگاه براساس روزهای هفته

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۵۳



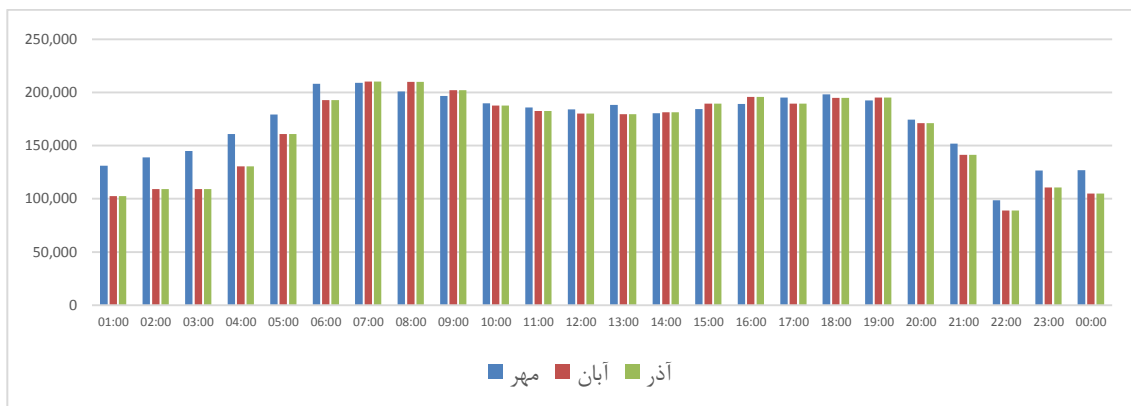
شکل ۱۵- حجم جستجوی انجام شده بر روی سامانه‌ی گنج از داخل کشور براساس روزهای هفته

نتایج نشان می‌دهد که روز یکشنبه در بین سه ماهه‌ی سوم سال ۱۳۹۵ دارای بالاترین حجم جستجو هم به طور کلی و هم در داخل پژوهشگاه می‌باشد. اما روزهای یکشنبه و سه‌شنبه به ترتیب روزهای پر جستجو در بین جستجوگران داخل پژوهشگاه هستند و روز سه‌شنبه روزی است که اکثر کاربران داخل کشور در حال جستجو بر روی سامانه‌ی گنج هستند.

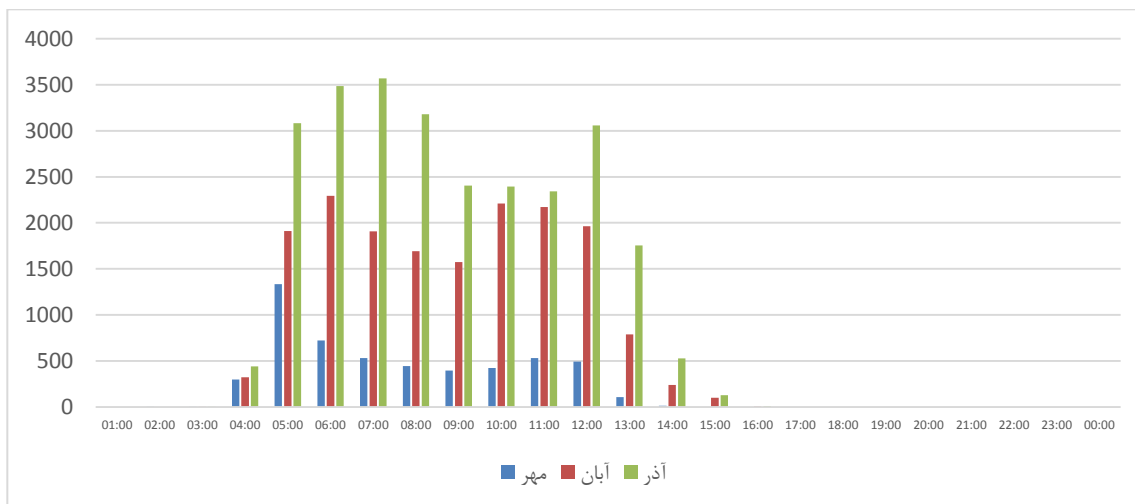
در تحلیل دیگری که بر روی داده‌های لاگ کاربران سامانه‌ی گنج صورت گرفت، حجم جستجوها براساس

ساعات شبانه روز بدست آمد.

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۵۴



شکل ۱۶- حجم کل جستجوها براساس ساعات شبانه‌روز به تفکیک ماه

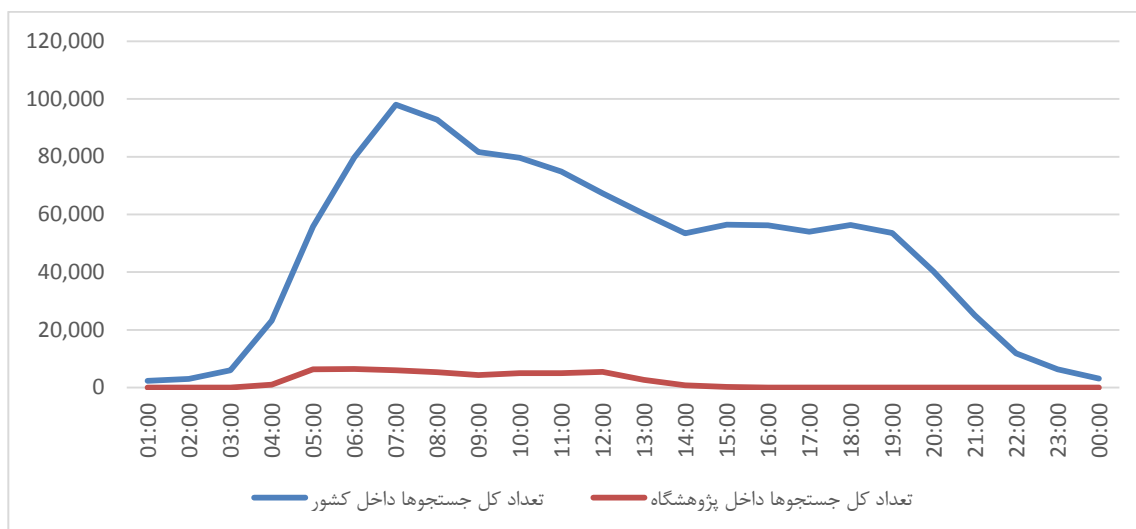


شکل ۱۷- حجم جستجوهای داخل پژوهشگاه براساس ساعات شبانه‌روز

نتایج نمودار نشان می‌دهد که بیشترین حجم جستجوها در بین ساعات ۷ الی ۹ صبح و ۵ الی ۷ بعدازظهر است. در زمانی که

افراد از داخل پژوهشگاه جستجو را انجام می‌دهند، زمان پر ترافیک جستجو در ساعت ۶ صبح و ۱۲ ظهر است.

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۵۵



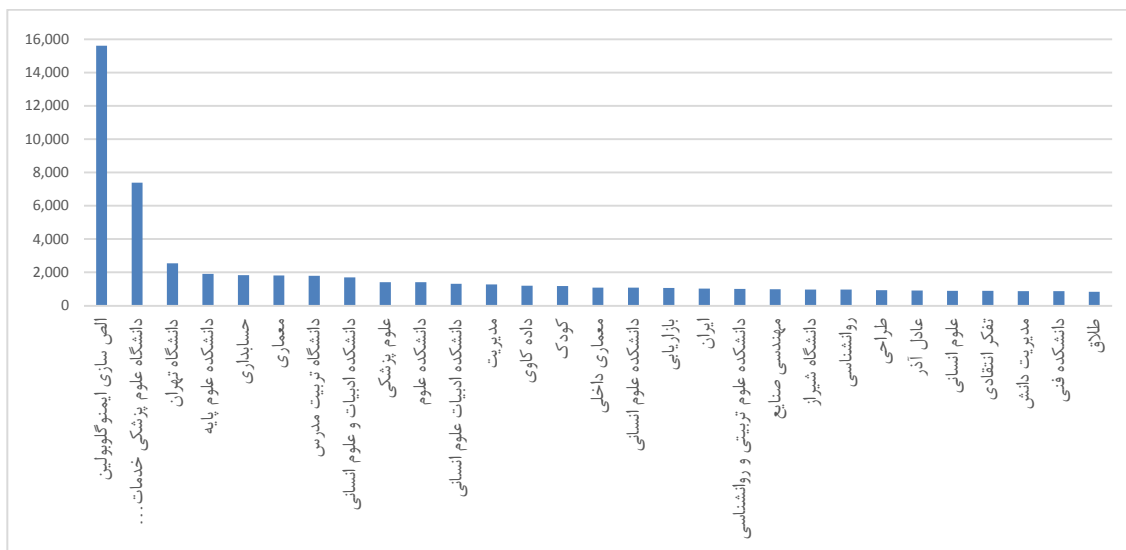
شکل ۱۸- مقایسه جستجوهای داخل پژوهشگاه و داخل کشور براساس ساعات شبانه‌روز

نتایج شکل ۱۸ نشان می‌دهد افراد در کشور در بازه ساعت بین ۷ الی ۹ صبح و همچنین ۷ عصر دارای دو قله با بیشترین میزان جستجو هستند.

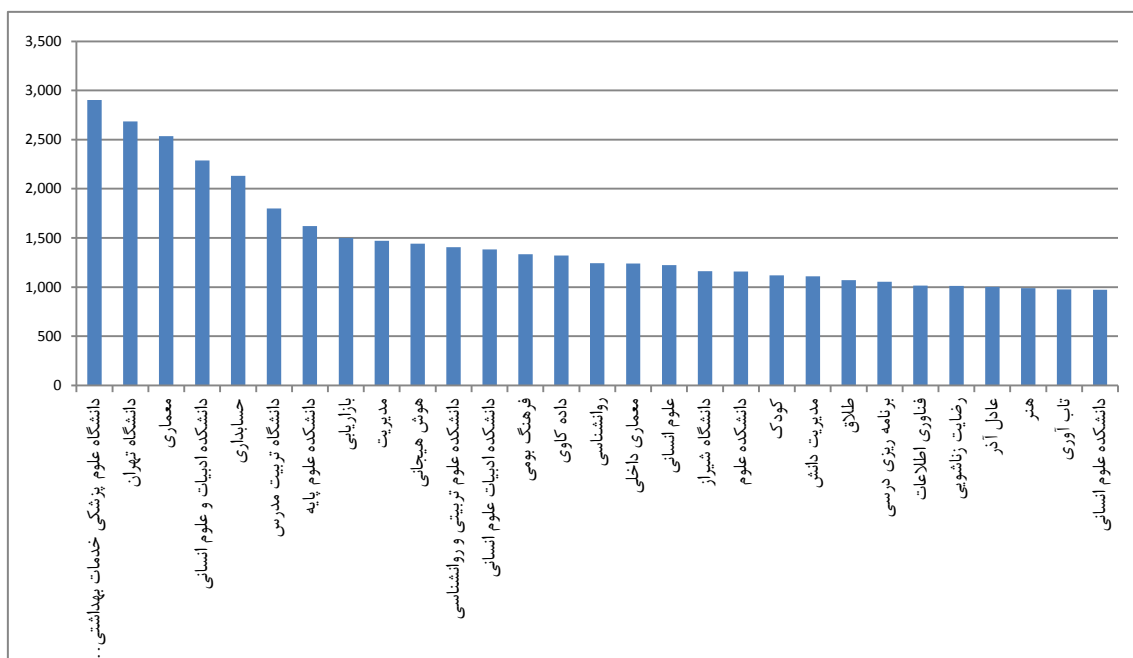
۳-۴- تحلیل داده‌ها براساس کلیدواژه‌ها

در تحلیل دیگری، بیشترین واژه‌های مورد جستجو در سامانه‌ی گنج نیز به تفکیک ماه بدست آمد. با توجه به اینکه در نگاه اولیه بر روی نمودارهای فوق الذکر، اکثر عبارات مورد جستجو مربوط به مکان‌ها هستند. تصمیم گرفته شد ۱۰۰ واژه‌ی پرتکرار در هر ماه به صورت جداگانه بررسی شود و برای هر کدام از آن‌ها یک برچسب متناسب با موضوع در نظر گرفته شود. این برچسب‌ها که شامل اماکن، روانشناسی، علوم اجتماعی، تجارت، مدیریت، فنی مهندسی، فناوری اطلاعات، فرهنگ، هنر، پزشکی، علوم تربیتی، حسابداری، علوم انسانی، ادبیات، افراد، محیط زیست، حقوق، سیاست و غیره است، به عبارت جستجو نسبت داده شدند.

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۵۶

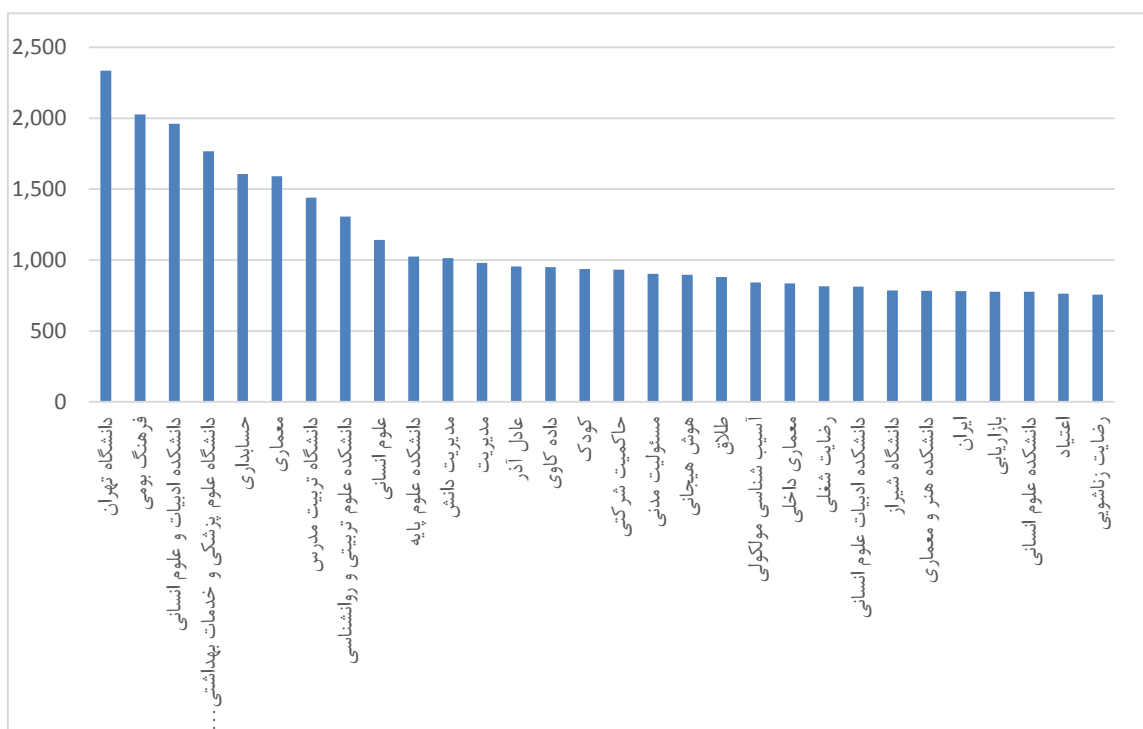


شکل ۱۹- بیشترین واژه‌های مورد جستجو در سامانه‌ی گنج در مهرماه

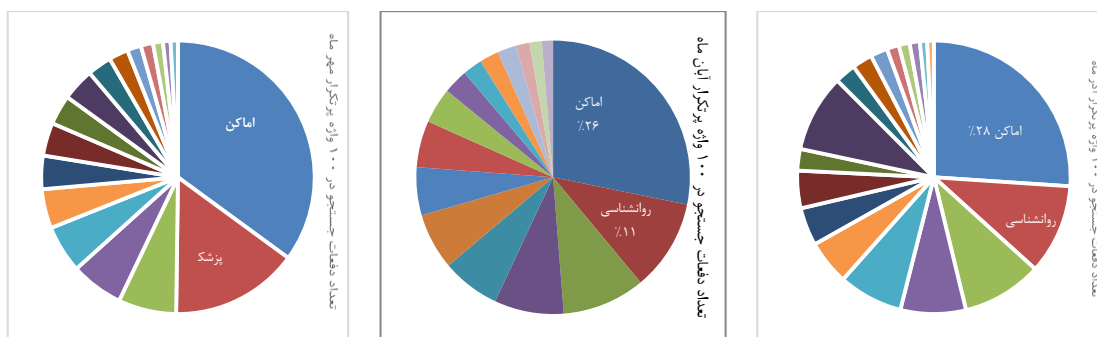


شکل ۲۰- بیشترین واژه‌های مورد جستجو در سامانه‌ی گنج در آبان ماه

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۵۷



شکل ۲۱- بیشترین واژه‌های مورد جستجو در سامانه‌ی گنج در آذر ماه



شکل ۲۲- سهم جستجوی لغات در بین ۱۰۰ واژه پرتکرار در جستجوها

همان‌طور که شکل‌های بالا نشان می‌دهند اکثر جستجوهای انجام شده عبارت‌هایی هستند که به مکان‌ها مربوط می‌شوند.

اگر امکان را در نظر نگرفته و کلیدواژه‌های پرتکرار را در نظر گرفته شود، لیست آن‌ها در جدول ۴ قابل مشاهده است.

جدول ۴- لیست کلیدواژه‌های پرتکرار در رشته‌های مختلف در کل جستجوها به همراه تعداد دفعات جستجو

• هوش هیجانی (۳۱۸۵) • روانشناسی (۲۸۳۶) • رضایت زناشویی (۲۵۳۷) • رضایت شغلی (۲۵۳۳) • افسردگی (۲۳۰۸)	روانشناسی (۲۸۰۱۷)
• خالص‌سازی ایمنوگلوبولین (۱۵۶۱۶) • علوم پزشکی (۲۳۸۳) • اوتیسم (۱۷۸۰)	پزشکی (۱۹۷۷۹)
• معماری (۵۹۷۴) • معماری داخلی (۳۲۲۸) • زنجیره تامین (۱۹۳۱) • طراحی معماری (۱۷۰۸) • مهندسی صنایع (۱۷۰۱)	فنی مهندسی (۱۶۱۲۷)
• طلاق (۲۸۰۹) • سرمایه اجتماعی (۲۰۸۰) • اعتیاد (۲۰۱۶) • اخلاق حرفه‌ای (۱۹۷۸) • سرمایه فکری (۱۹۶۳)	علوم اجتماعی (۱۵۹۸۹)
• مدیریت (۳۷۴۸) • مدیریت دانش (۳۰۳۱) • مدیریت مالی (۱۷۷۰) • تعهد سازمانی (۱۷۶۹) • کارآفرینی (۱۷۶۹)	مدیریت (۱۵۳۰۸)
• داده کاوی (۳۴۹۵) • فناوری اطلاعات (۲۴۳۹) • تحلیل محتوا (۲۲۴۰) • فضای مجازی (۱۸۴۷) • پردازش تصویر (۱۸۰۴)	فناوری اطلاعات (۱۵۱۳۱)
• بازاریابی (۳۳۵۵) • تولید (۲۰۱۰) • توسعه پایدار (۱۹۹۸) • اقتصاد (۱۹۸۲) • اقتصاد مقاومتی (۱۷۹۵)	تجارت (۱۲۷۷۹)
• تصویرسازی (۲۳۷۱) • هنر (۲۳۶۴) • سینما (۱۷۸۱) • گرافیک (۱۷۷۸)	هنر (۹۹۰۷)
• علوم انسانی (۳۲۶۸) • مسئولیت مدنی (۲۳۳۹) • برنامه ریزی درسی (۲۲۰۷)	علوم انسانی (۷۸۱۴)

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۵۹

در حالت کلی لیست کلیدواژه‌های پرتکرار ارائه شد، اگر تنها IP های داخل پژوهشگاه مد نظر قرار گیرد، لیست کلیدواژه- های مورد جستجوی کاربران تا حدی متفاوت است (جدول ۵).

جدول ۵- لیست کلیدواژه‌های پرتکرار در رشته‌های مختلف در جستجوهای داخل پژوهشگاه

روانشناسی	• هوش چندگانه (۱۳) • سلامت روان تاب آوری (۱۱)
پزشکی	• بیماری های کودکان (۱۲) • حسی-حرکتی (۱۲)
فنی مهندسی	• معماری (۱۲۳) • برق (۳۰) • فرآیند طراحی معماری (۱۴) • طراحی رویکرد مصرف انرژی (۱۳)
علوم اجتماعی	• تغییر جنسیت حقوق حقوقی (۱۲) • فسخ نکاح (۱۱)
مدیریت	• مدیریت دانش (۹۱) • نقشه دانش (۲۶) • برنامه استراتژیک & مدیریت دانش (۱۷)
فناوری اطلاعات	• آموزش مجازی (۶۳) • آموزش الکترونیکی (۲۵) • امکان سنجی آموزش مجازی (۱۹) • دانشگاه مجازی (۱۷) • سایر (۱۷)
تجارت	• اقتصاد مقاومتی (۳۴)
هنر	• فرهنگ & رسانه (۱۹)
علوم انسانی	• علم اطلاعات و دانش شناسی (۳۶) • فلسفه (۲۳) • مثنوی (کتاب) (۲۲) • رفتار اطلاع یابی (۱۴)

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۶۰

لیست کلیدواژه‌های پرتکرار که از سراسر کشور مورد جستجوی کاربران بوده است در جدول ۶ آمده است.

جدول ۶- لیست کلیدواژه‌های پرتکرار در رشته‌های مختلف در جستجوهای داخل کشور

• طلاق (۴۹۶) • روانشناسی (۴۶۷) • هوش هیجانی (۴۶۲) • افسردگی (۴۵۲)	روانشناسی
• اوتیسم (۳۹۵)	پزشکی
• معماری (۱۰۱۵) • مهندسی صنایع (۵۳۵) • معماری داخلی (۴۹۷)	فنی مهندسی
• اعتیاد (۳۵۹)	علوم اجتماعی
• مدیریت (۷۲۹) • مدیریت دانش (۶۷۴) • مدیریت مالی (۴۸۱)	مدیریت
• داده‌کاوی (۷۰۵) • فناوری اطلاعات (۶۱۸)	فناوری اطلاعات
• بازاریابی (۶۹۱) • زنجیره تامین (۶۷۶) • اقتصاد مقاومتی (۵۲۸) • اقتصاد (۴۷۳)	تجارت
• تصویرسازی (۴۵۵) • سینما (۴۰۲)	هنر
• حسابداری (۱۲۲۴) • علوم انسانی (۵۹۲) • فرهنگ بومی (۵۲۰)	علوم انسانی

مقایسه نتایج جدول ۴، ۵ و ۶ نشان می‌دهد افرادی که از داخل پژوهشگاه جستجو کرده‌اند، هدفمندتر و دقیقتر عبارات خود

را برای جستجو انتخاب کرده‌اند.

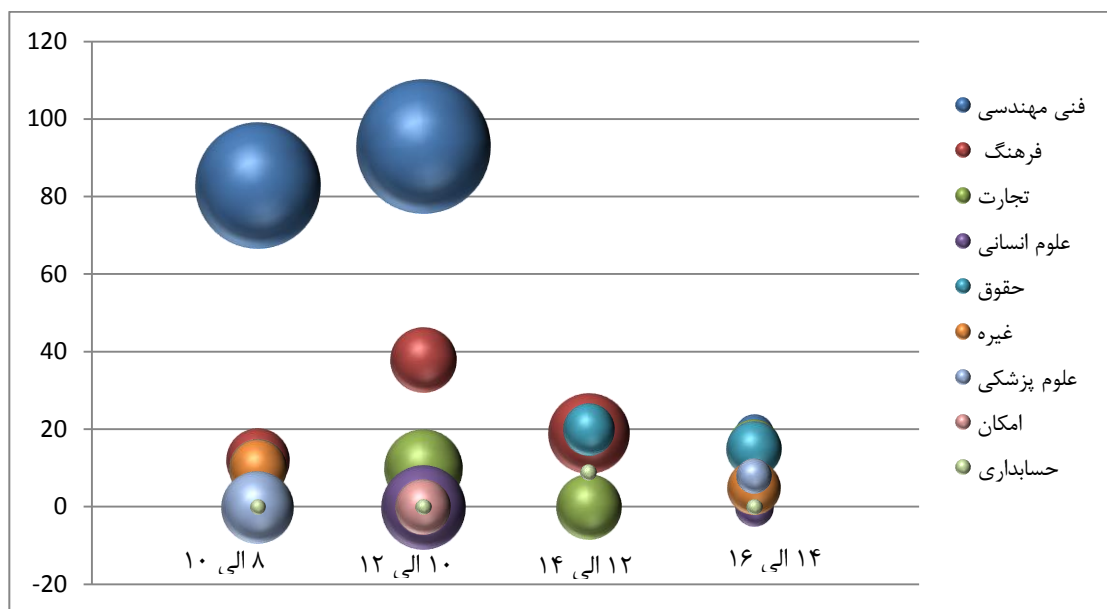
۴-۴- تحلیل داده‌ها براساس کلیدواژه‌ها و زمینه‌های علمی

همچنین کلمات پرتکرار برحسب زمینه‌های مختلف جستجو در ساعات مختلف روز مورد بررسی قرار گرفت و نتایج آن در جدول ۷ و شکل ۲۳ و ۲۴ قابل مشاهده است. به عنوان مثال افراد علاقه‌مند به موضوعات فنی مهندسی و فناوری اطلاعات بیشتر در صبح و ساعات ۸ الی ۱۲ جستجوهای خود را انجام می‌دهند. این در حالی است که موضوعات مربوط به علوم انسانی بیشتر در ساعات ۱۲ الی ۱۶ مورد جستجو است.

جدول ۷- آمار مربوط به جستجوی کلمات برحسب ساعات مختلف جستجو

موضوعات مورد جستجو	هشت الی ده صبح	ده تا دوازده صبح	دوازده الی چهارده بعدازظهر	چهارده الی شانزده بعدازظهر
فنی مهندسی	۸۳	۹۳	۱۹	۱۹
فناوری اطلاعات	۶۷	۷۶	۰	۶
فرهنگ	۱۲	۳۸	۱۹	۶
مدیریت	۱۷	۱۸	۲۸	۰
تجارت	۳۱	۱۰	۰	۱۶
روانشناسی	۰	۲۶	۱۸	۱۱
علوم انسانی	۰	۰	۴۹	۰
علوم اجتماعی	۱۰	۳۰	۰	۶
حقوق	۱۰	۰	۲۰	۱۵
افراد	۱۴	۰	۱۱	۱۳
غیره	۱۰	۰	۲۱	۵
مذهب	۱۴	۹	۰	۱۲
علوم پزشکی	۰	۱۲	۱۰	۸
هنر	۲۲	۰	۰	۵
امکان	۰	۰	۱۴	۰
علوم پایه	۰	۱۳	۰	۰
حسابداری	۰	۰	۹	۰

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۶۲



شکل ۲۳- آمار مربوط به جستجو بر حسب ساعات مختلف جستجو بر اساس زمینه علمی

ساعت ۸ الی ۱۰ صبح		ساعت ۱۲ الی ۲ بعد از ظهر		ساعت ۱۰ الی ۱۲ ظهر		ساعت ۲ الی ۴ بعد از ظهر	
کلیدواژه جستجو شده	تعداد دفعات جستجو	کلیدواژه جستجو شده	تعداد دفعات جستجو	کلیدواژه جستجو شده	تعداد دفعات جستجو	کلیدواژه جستجو شده	تعداد دفعات جستجو
معماری	۵۷	فرهنگ و رسانه	۱۹	معماری	۶۵	محمود نقیب زاده	۱۳
آموزش الکترونیکی	۱۸	علم اطلاعات و دانش‌شناسی کتابداری و ام	۱۸	آموزش مجازی	۳۴	بسته بندی دارو	۸
برنامه استراتژیک و مدیریت دانش	۱۷	ایران	۱۴	نقشه دانش	۱۸	حفظ قرآن و زندگی و روایات	۷
انضاری مجتبی	۱۴	فراتحلیل	۱۳	تلی زاگویی حل معادلات دیفران	۱۷	آنتولوژی	۷
آموزش مجازی	۱۴	عبدالوهاب بیانی	۱۱	امکان سنجی آموزش	۱۵	آر دی ای	۶
دین "فلسفه اسلامی"	۱۴	نظریه و اطلاعات دانش	۱۱	امکان سنجی آموزش مجازی	۱۵	شبکه مجازی و تبلیغات	۶
فرایند طراحی معماری	۱۴	نظریه دانش	۱۱	مثنوی (کتاب)	۱۵	شبیه سازی آشکارساز	۶
دانشگاه مجازی	۱۳	چگونگی انتخاب سیستم ارتباطات اطلاعات مناد	۱۱	رفتار اطلاع یابی	۱۴	عوامل رفتار خریدار	۶
زبان و شعر و اخوان	۱۳	حقوق اداری	۱۱	تفجار اطلاعات	۱۳	دولت و حمایت و مصرف‌کننده	۶
امکان سنجی مجازی	۱۲	حسی حرکتی	۱۰	فیزیک	۱۳	مسئولیت مدنی و بیمه	۶
معنا و معماری	۱۲	تولید پراکنده قدرت فازی	۱۰	رسانه و فرهنگ	۱۳	هوش هیجانی دانش آموز عادی/تیزهوش	۶
نقش گرافیک محیطی بر	۱۲	هوش چندگانه	۱۰	طراحی سالمندان	۱۲	هنر دیجیتال معاصر	۵
اقتصاد مقاومتی	۱۱	بازی یازدها زنجیره تأمین/تامین	۹	رویکرد صرفه جویی انرژی پایدار	۱۱	حقوقی/حقوق کودک آزاری	۵
پیامد نتیجه اثر اقتصادی گزارش گزارشگری مالی مؤ	۱۰	تخصص حسابرس و کیفیت حسابره	۹	سلامت روان تاب آوری	۱۱	قطعنامه و شورای امنیت و حقوق بشر	۵
تبعیض نژاد/نژادی حقوق/حقوقی	۱۰	مشارکت اجتماعی و ارتباطات سازه	۹	نظامی/نظامی گری ایران	۱۱	مالکیت فکری و حقوق و مفهوم	۵
حمايه بهار	۱۰	عدم تقارن اطلاعات/اطلاعاتی تقسیم سود سا	۹	سرقت علمی	۱۰	مجموعه	۵
اقتصاد مقاومتی	۱۰	دلیل/داده مسئولیت مدنی فقه	۹	گری مالی سرمایه گذاری منابع	۱۰	ارکتنان "رضایت شفاهی" فرهنگ سازمانی	۵
سایبر	۱۰	ساخت سفارش/آبشار	۸	تاریخ محلات	۱۰	کیفیت حسابرسی "مدیریت سود"	۵

شکل ۲۴- آمار مربوط به جستجوی کلمات بر حسب ساعات مختلف جستجو

با توجه به اینکه در سه ماهه پاییز ۹۵، ۵۶ روز کاری وجود داشته است و تعداد کل کلید واژه‌های جستجو ۴۸۸۹۹ عدد

بوده است، بنابراین به‌طور متوسط در هر روز ۸۷۳ کلیدواژه جستجو شده است و در هر ساعت تنها ۱۰۹ واژه در سامانه گنج

به‌عنوان کلیدواژه مورد جستجو قرار گرفته است. همچنین با توجه به آمار تعداد حضور کاربران که روزانه به‌طور متوسط ۱۸۰ تن ثبت شده است، کاربران با متوسط ۴٫۸۵ واژه به نتیجه رسیده‌اند و یا جستجوی خود را متوقف نموده‌اند. بدیهی است برای جستجوی حوزه پژوهشی در نگارش پروپزال، پایان‌نامه و یا رساله اکتفا نمودن به حداکثر پنج واژه نمی‌تواند پژوهش مدنظر را جامع نماید.

همان‌طور که در جدول ۷ و شکل ۲۳ ملاحظه می‌گردد کاربران علاقمند به حوزه فنی مهندسی بیشتر در ابتدای روز به پژوهشگاه مراجعه می‌نمایند، در حالی که علاقمندان به حقوق در ساعات بعد از ظهر مراجعه می‌نمایند. کلیدواژه تجارت در کل روز تقریباً به‌طور مساوی مراجعه‌کننده خود را دارد و علاقمندان فرهنگ بیشتر ساعات میانی را ترجیح می‌دهند. به‌طور کلی کاربران علوم انسانی ساعات بعدازظهر را برای جستجوی خود مطلوب‌تر می‌دانند. بنابراین به نظر می‌رسد که می‌توان تبلیغات برنامه‌های آموزشی را به تفکیک حوزه‌های کاربران در طول روز بخش‌بندی نمود به‌طوری که تبلیغات نیمه اول روز بیشتر معطوف به حوزه فنی مهندسی باشد و حوزه علوم انسانی در بعدازظهر انجام گیرد.

همچنین با توجه به اینکه اماکن بیشترین جستجو را به خود اختصاص داده است، می‌توان با آموزش کاربران جستجوی آن‌ها را طوری هدایت نمود که بتوانند به نحو مطلوب‌تری به تسلط نسبی در حوزه پژوهش مورد علاقه خود دست یابند و استادان و پژوهشگران این حوزه را با سرعت بیشتری شناسایی نمایند.

۵-۴- تحلیل داده‌ها براساس خوشه‌بندی کاربران

یکی از اهدافی که در این پژوهش مطرح بود، خوشه‌بندی کاربران از لحاظ عبارات جستجو و زمان جستجو بود. اما نکته‌ای که در داده‌های واقعی سامانه‌ی گنج وجود داشت، عدم قابلیت تفکیک افراد براساس id و یا نشست بود. زیرا تمام اطلاعات مورد نیاز تفکیک کاربران لاگ نشده بود. نمونه‌ای از این مشکل در شکل ۲۵ و ۲۶ قابل مشاهده است. به عنوان مثال کلمات کلیدی "موشک ایران"، "موشکی ایران"، "موشکی ایران" توسط یک کاربر و کلمات "مدل چابکی فناوری اطلاعات" توسط فرد دیگری جستجو شده‌اند اما از آنجایی که برای تمامی افراد یک user_id در نظر گرفته شده است. امکان تفکیک وجود ندارد.

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۶۴

Query	User_ID	Created_at	Weekday	Hour	Number	IP
نوآوری تکنولوژی	375827	12/3/2016 6:47	1	6	9	10.19.157.202
فعالیت موشکی ایران بین الملل	375827	12/3/2016 6:49	1	6	0	10.19.157.202
موشک ایران بین الملل	375827	12/3/2016 6:49	1	6	4	10.19.157.202
نوآوری تکنولوژی صادرات	375827	12/3/2016 6:50	1	6	24	10.19.157.202
موشک ایران	375827	12/3/2016 6:49	1	6	89	10.19.157.202
موشک ایران	375827	12/3/2016 6:50	1	6	0	10.19.157.202
رابطه نوآوری تکنولوژی صادرات	375827	12/3/2016 6:50	1	6	11	10.19.157.202
موشک ایران	375827	12/3/2016 6:50	1	6	89	10.19.157.202
مدلی مدل چابکی چابک "فناوری اطلاعات" کسب و کار	375827	12/3/2016 6:49	1	6	18	10.19.157.202
موشکی ایران	375827	12/3/2016 6:49	1	6	4	10.19.157.202
مدلی مدل چابکی چابک "فناوری اطلاعات" کسب کار	375827	12/3/2016 6:49	1	6	17	10.19.157.202
موشکی موشک بین الملل ایران	375827	12/3/2016 6:50	1	6	0	10.19.157.202
موشکی موشک بین الملل ایران	375827	12/3/2016 6:51	1	6	0	10.19.157.202
موشکی موشک بین الملل ایران	375827	12/3/2016 6:51	1	6	0	10.19.157.202
مدلی مدل چابکی چابک "فناوری اطلاعات" کسب کار	375827	12/3/2016 6:51	1	6	17	10.19.157.202
کمیته حسابرسی داخلی پاسخ دهی مالی بانک	375827	12/19/2016 13:15	3	13	0	10.19.157.202
کمیته حسابرسی داخلی پاسخ مالی	375827	12/19/2016 13:16	3	13	5	10.19.157.202
کمیته حسابرسی داخلی پاسخ مالی	375827	12/19/2016 13:16	3	13	3	10.19.157.202
کمیته حسابرسی داخلی پاسخگویی مالی	375827	12/19/2016 13:17	3	13	3	10.19.157.202
کمیته حسابرسی داخلی پاسخگویی مالی	375827	12/19/2016 13:17	3	13	3	10.19.157.202
نظریه علت تعهد حقوق	375827	12/21/2016 11:47	5	11	17	10.19.157.202
نظریه علت تعهد حقوق	375827	12/21/2016 11:48	5	11	17	10.19.157.202
علت تعهد حقوق	375827	12/21/2016 11:49	5	11	78	10.19.157.202
علت تعهد حقوق گامن لا	375827	12/21/2016 11:49	5	11	5	10.19.157.202
نظریه علت حقوق گامن لا	375827	12/21/2016 11:49	5	11	3	10.19.157.202

شکل ۲۵- نمونه ای از لاگ های جستجو با id یکسان کاربران

Query	User_ID	Created_at	Number	IP	Session_ID
سرمایه اجتماعی عملکرد	375827	09/24/2016 04:53	51	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
آلبانو اینولین	375827	09/24/2016 04:54	2	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد استراتژیک	375827	09/24/2016 04:55	0	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد استراتژیک	375827	09/24/2016 04:56	0	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد تکنولوژی	375827	09/24/2016 04:56	0	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد عدم قطعیت تکنولوژی	375827	09/24/2016 04:57	1	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد تکنولوژی	375827	09/24/2016 04:58	58	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد قطعیت تکنولوژی	375827	09/24/2016 04:58	1	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد فناوری تکنولوژی	375827	09/24/2016 04:58	129	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد قطعیت تکنولوژی	375827	09/24/2016 04:58	1	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد فناوری تکنولوژی	375827	09/24/2016 04:59	129	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد فناوری تکنولوژی	375827	09/24/2016 04:59	2	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
سرمایه اجتماعی عملکرد فناوری تکنولوژی	375827	09/24/2016 05:00	119	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
شنا بیش فعالی	375827	09/24/2016 05:00	1	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
شنا تقص توجه	375827	09/24/2016 05:01	6	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
شنا بیش فعالی	375827	09/24/2016 05:02	1	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
شنا آرام سازی	375827	09/24/2016 05:04	3	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
adhd	375827	09/24/2016 05:04	0	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
اینولین توت فرنگی	375827	09/24/2016 05:06	0	10.19.157.202	25b79f29ecf6265c6d24837396361fc4
اینولین توت فرنگی	375827	09/24/2016 05:06	0	10.19.157.202	25b79f29ecf6265c6d24837396361fc4

شکل ۲۶- نمونه ای از لاگ های جستجو با id یکسان نشست

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۶۵

همانطور که شکل ۲۵ نشان می‌دهد session_id یکسانی در نظر گرفته شده است و تمام جستارهای مورد نظر روی این نشست فهرست شده‌اند، با توجه به اینکه رنگ‌های مختلف در شکل نماینگر موضوعات مختلف هستند و واضح است که توسط کاربران مختلف جستجو شده‌اند اما امکان تفکیک آن‌ها وجود ندارد. با توجه به مشکلات عدیده‌ای که در لاگ وجود داشت و امکان استخراج اطلاعات مفید جهت خوشه‌بندی موجود نبود، یک مجموعه داده را که شامل ip داخل پژوهشگاه می‌شوند را جدا کردیم و آن‌ها را برچسب‌گذاری کردیم. قابل ذکر است که به دلیل حجم بالای داده‌ها، فقط ۱۱۰۰ تا از کوئری‌ها برحسب موضوع جستجو توسط افراد خبره برچسب‌گذاری شدند. فهرست برچسب موضوعی و تعداد آن‌ها در جدول زیر آمده است:

جدول ۸- آمار تعداد جستارهای برچسب‌گذاری شده توسط خبره

برچسب موضوعی	تعداد
تجارت	۲۳۹
روانشناسی	۱۶۴
حقوق	۱۲۳
مدیریت	۱۱۸
علوم اجتماعی	۹۳
فنی مهندسی	۶۳
مذهب	۵۷
فناوری اطلاعات	۵۶
حسابداری	۵۲
سیاست	۳۱
غیره	۲۷
اماکن	۲۵
هنر	۲۳
علوم تربیتی	۱۷
علوم پزشکی	۶
محیط زیست	۴
افراد	۲

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۶۶

پس از این مرحله، مجموعه داده آماده شده و عملیات خوشه‌بندی را بر روی ۱۱۰۰ داده با ویژگی‌های شماره روز هفته، ساعت جستجو و برجسب موضوعی که به آن اختصاص داده شده بود برای حالت دو خوشه انجام گرفت. الگوریتم مورد استفاده k-means است. نتایج در شکل ۲۷ ارائه شده است. منظور از درصد این است که چه درصدی از داده‌های مربوط به یک برجسب در یک خوشه قرار گرفته‌اند. به عنوان مثال عدد ۱۰۰ در مورد موضوع فناوری اطلاعات نشان می‌دهد صددرصد و یا درواقع تمام موضوعات با عنوان فناوری اطلاعات در خوشه اول قرار دارند.

خوشه اول	فناوری اطلاعات	فنی مهندسی	حسابداری	علوم تربیتی	علوم پزشکی	افراد	اماکن	محیط زیست	هنر	غیره
درصد	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰
خوشه دوم	تجارت	حقوق	روانشناسی	سیاست	مدیریت	علوم اجتماعی	مذهب			
درصد	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰			

شکل ۲۷- نتایج خوشه‌بندی برای دو خوشه

الگوریتم k-means برای حالت سه و چهار خوشه نیز اجرا شد که نتایج آن در شکل ۲۶ و ۲۷ آمده است.

خوشه اول	مذهب	علوم اجتماعی	حقوق	تجارت	مدیریت	روانشناسی				
درصد	۵۷.۸۹	۵۹.۱۴	۵۲.۰۳	۴۸.۵۴	۷۹.۶۶	۷۷.۴۴				
خوشه دوم	اماکن	فنی مهندسی	سیاست	علوم پزشکی	هنر	علوم تربیتی	غیره	محیط زیست	افراد	حسابداری
درصد	۰.۹۶	۰.۷۸	۰.۲۳	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰
خوشه سوم	روانشناسی	مدیریت	تجارت	حقوق	علوم اجتماعی	مذهب	سیاست	فنی مهندسی	اماکن	
درصد	۲۲.۵۶	۲۰.۳۴	۵۱.۴۶	۴۴.۷۲	۴۰.۸۶	۴۲.۱۱	۷۷.۴۲	۲۲.۲۲	۴.۰۰	

شکل ۲۸- نتایج خوشه‌بندی برای سه خوشه

خوشه اول	اماکن	فنی مهندسی	علوم پزشکی	هنر	علوم تربیتی	غیره	محیط زیست	افراد	حسابداری	فناوری اطلاعات
درصد	۰.۸۴	۰.۴۶	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰
خوشه دوم	علوم اجتماعی	حقوق	تجارت	مدیریت	روانشناسی					
درصد	۰.۰۶	۰.۲۳	۰.۳۶	۰.۲۰	۰.۲۳					
خوشه سوم	روانشناسی	مدیریت	تجارت	حقوق	علوم اجتماعی	مذهب				
درصد	۰.۷۷	۰.۸۰	۰.۴۹	۰.۴۹	۰.۵۶	۰.۳۵				
خوشه چهارم	تجارت	حقوق	علوم اجتماعی	مذهب	سیاست	فنی مهندسی	اماکن			
درصد	۰.۱۶	۰.۲۸	۰.۳۸	۰.۶۵	۱۰۰.۰۰	۰.۵۴	۰.۱۶			

شکل ۲۹- نتایج خوشه‌بندی برای چهار خوشه

همانطور که نتایج نشان می‌دهد برخی از موضوعات مانند فناوری اطلاعات، حسابداری و فنی مهندسی اغلب در یک خوشه و موضوعاتی مانند تجارت و حقوق اغلب در یک خوشه قرار می‌گیرند.

با توجه به اینکه موضوعاتی مانند محیط زیست، افراد، هنر و علوم پزشکی تعداد کمی از حجم داده‌های منتخب را تشکیل می‌دهند، بر آن شدیم آن‌ها را حذف کرده و دسته‌هایی را برای مجموعه داده انتخاب کنیم که حجم بیشتری از برچسب موضوعی را شامل می‌شوند. آمار مربوط به تعداد جستارها در هر زمینه‌ی موضوعی در مجموعه داده‌ی جدید آمده است (جدول ۹):

جدول ۹- آمار تعداد جستارهای برچسب‌گذاری شده توسط خبره پس از حذف برچسب‌های موضوعی کم تکرار

برچسب	تعداد
روانشناسی	۱۶۴
مدیریت	۱۱۸
تجارت	۲۳۹
حقوق	۱۲۳
علوم اجتماعی	۹۳
مذهب	۵۷
سیاست	۳۱
فنی مهندسی	۶۳
فناوری اطلاعات	۵۶
حسابداری	۵۲

با مجموعه داده‌ی جدید عملیات خوشه‌بندی با استفاده از الگوریتم k-means مجدداً انجام شد و نتایج آن در شکل ۳۰ آمده است:

خوشه اول	فناوری اطلاعات	حسابداری	فنی مهندسی	سیاست	مذهب
درصد	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۰.۸۱
خوشه دوم	روانشناسی	حقوق	تجارت	مذهب	مدیریت
درصد	۱۰۰.۰۰	۱۰۰.۰۰	۱۰۰.۰۰	۱۹.۳۰	۱۰۰.۰۰

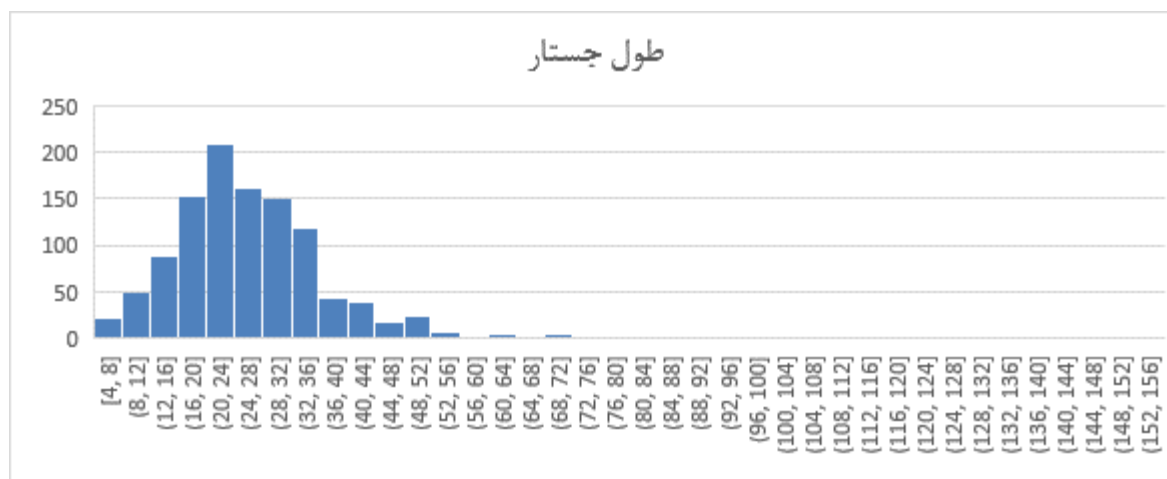
شکل ۳۰- نتایج خوشه‌بندی برای دو خوشه با حذف داده‌های کم تکرار

همانطور که نتایج شکل ۳۰ نشان می‌دهد، همچنان دسته‌ی موضوعی فناوری اطلاعات، حسابداری و فنی مهندسی در یک خوشه و موضوعات روانشناسی، حقوق و تجارت در خوشه‌ی دیگری قرار می‌گیرند.

در ادامه جهت تفکیک بهتر داده‌ها در خوشه‌های مجزا تصمیم گرفته شد، ویژگی طول جستار و یا در واقع طول کلیدواژه مورد جستجو را نیز به مجموعه داده‌ها اضافه کنیم. میانگین طول کلیدواژه جستجو در هر زمینه‌ی علمی به تفکیک در جدول ۱۰ آمده است. همچنین در شکل ۳۱ فراوانی طول کلیدواژه‌های جستجو آمده است.

جدول ۱۰- میانگین طول کلیدواژه جستجو به تفکیک زمینه علمی

زمینه علمی	میانگین طول کلیدواژه جستجو
فناوری اطلاعات	۳۱,۹۵
مدیریت	۳۱,۰۲
فنی مهندسی	۳۱,۰۲
هنر	۳۰,۲۲
حسابداری	۲۹,۳۳
علوم اجتماعی	۲۸,۹۸
روانشناسی	۲۷,۷۸
سیاست	۲۷,۵۸
تجارت	۲۷,۰۶
علوم تربیتی	۲۴,۷۱
علوم پزشکی	۲۴,۵۰
غیره	۲۲,۲۲
مذهب	۲۱,۷۴
حقوق	۲۰,۱۰
محیط زیست	۱۳,۷۵
اماکن	۱۳,۱۲
افراد	۸,۰۰



شکل ۳۱- توزیع کلیدواژه‌های جستجو شده بر حسب طول

همانطور که شکل ۳۱ نشان می‌دهد اکثر عبارات مورد جستجو بین ۲۰ تا ۲۴ کاراکتر دارای طول هستند. اما کلماتی با طول

۱۵۴ کاراکتر هم در داده‌ها مشاهده شدند. در نهایت، ویژگی طول عبارات مورد جستجو به مجموعه داده‌ها اضافه گردید

(شکل ۳۲).

تحلیل رفتار جستجوی اطلاعات پژوهشگران در موتور جستجوی سامانه‌ی گنج □ ۷۰

B	C	D	E	F
Weekday	Hour	Number	Len	Tag
1	6	7	28	روانشناسی
2	7	19	30	مدیریت
2	7	1	33	تجارت
2	7	3	28	تجارت
2	7	0	25	تجارت
5	11	13	24	تجارت
5	11	13	24	تجارت
5	11	11	26	تجارت
5	11	11	26	تجارت
1	6	2	26	حقوق
1	6	124	18	حقوق
1	6	7	18	حقوق
1	6	1	24	تجارت
1	6	1	21	تجارت
1	6	1	23	روانشناسی
5	11	21	36	تجارت
5	11	9	32	علوم اجتماعی
5	11	21	36	روانشناسی
4	6	28	9	حقوق
4	6	1	15	حقوق
4	6	9	20	حقوق
4	6	1	22	حقوق
4	6	1	22	حقوق

شکل ۳۲- مجموعه داده برچسب گذاری شده توسط خبره براساس ویژگی‌های مستخرج از لاگ

داده‌ها براساس مقادیر بیشینه و کمینه در هر بعد از ویژگی‌ها نرمالسازی شدند و بر روی مجموعه داده بدست آمده عمل

خوشه‌بندی برای ۲، ۳ و ۴ خوشه انجام شد که نتایج آن در زیر قابل مشاهده است.

جدول ۱۱- نتایج خوشه‌بندی بر روی مجموعه داده‌ها شامل طول عبارات جستجو برای دو خوشه

		Weekday	Hour	Number	Len
Cluster 1	Avg	4.48	8.75	447.66	27.32
	STD	0.54	3.18	5492.07	14.1
Cluster 2	AVG	1.29	9.62	120.18	27.73
	STD	0.52	2.67	678.50	10.61

همانطور که نتایج خوشه‌بندی براساس میانگین و انحراف از معیار نشان می‌دهد یک خوشه افرادی هستند که

روزهای ابتدای هفته و ساعات نخستین روز را برای جستجو انتخاب می‌کنند و دفعات جستجوی آنها زیاد است. اما دسته

دیگر افرادی هستند که روزهای آخر هفته و ساعات نزدیک به ظهر را برای جستجو انتخاب می‌کنند و تعداد جستجوهای

آنها کمتر است.

جدول ۱۲- نتایج خوشه‌بندی بر روی مجموعه داده‌ها شامل طول عبارات جستجو برای سه خوشه

		Weekday	Hour	Number	Len
Cluster 1	Avg	1.28	11.49	77.73	27.17
	Std	0.57	1.46	362.3	10.8
Cluster 2	Avg	1.29	7.03	179.24	28.51
	STD	0.45	1.59	955.11	10.29
Cluster 3	Avg	4.47	8.75	447.67	27.32
	Std	0.53	3.18	5492.07	14.09

جدول ۱۳- نتایج خوشه‌بندی بر روی مجموعه داده‌ها شامل طول عبارات جستجو برای چهار خوشه

		Weekday	Hour	Number	Len
Cluster 1	Avg	4	5.54	886.83	28.10
	Std	0	1.19	8125.38	17.79
Cluster 2	Avg	4.72	11.55	100.67	27.76
	std	0.65	1.37	475.21	11.69
Cluster 3	Avg	1.39	6.32	213.11	28.40
	std	0.48	1.24	1073.07	10.44
Cluster 4	avg	1.15	10.97	67.98	26.77
	std	0.36	1.58	359.41	9.51

۶-۴- نتیجه‌گیری

در این بخش، نتایج بدست آمده از تحلیل لاگ‌های کاربران سامانه‌ی گنج در سه بخش تحلیل داده‌ها براساس زمان، کلیدواژه‌ها و کلید واژه‌ها و زمینه‌های علمی مورد بررسی قرار گرفت و نتایج بین جستجوهای داخل و خارج از پژوهشگاه تفاوت‌های قابل توجهی را گزارش کرد. همچنین نتایج نشان داد که بین زمینه‌های علمی مختلف و ساعات جستجو ارتباط وجود دارد.

فصل پنجم

جمع بندی و کارهای آتی

۱-۵- نتیجه‌گیری

در این پژوهش به تحلیل لاگ‌های کاربران سامانه‌ی گنج پرداخته شد و نتایج این تحلیل نشان می‌دهد که بیش از ۹۹٪ کاربران تمایل دارند تا از بیرون پژوهشگاه به جستجو بپردازند و کمتر از ۱٪ از آن‌ها به دلیل دسترسی به متن کل پایان‌نامه و رساله در تالار جستجوی پژوهشگاه مراجعه حضوری دارند. این امر خود بیانگر آنست که موتور جستجوی پایگاه گنج باید به نحو مطلوب‌تری عمل نماید تا با توجه به اینکه درصد بسیاری از پژوهشگران در تهران حضور دارند، بتواند تالار جستجو تعداد کاربر بیشتری را جذب نماید. همچنین با توجه به اینکه هر کاربر در مراجعه حضوری به طور متوسط پس از جستجوی پنج کلید واژه فرایند جستجوی خود را متوقف می‌نماید نیاز است که پایگاه گنج به نحو جذاب‌تری طراحی گردد تا پژوهشگر بتواند در حوزه پژوهش خود دانش بیشتری را کسب نماید. نتایج پژوهش بیانگر آنست که پژوهشگران حوزه علوم انسانی بیشتر علاقمند به حضور در ساعات بعد از ظهر می‌باشند و پژوهشگران فنی مهندسی بیشتر ساعات اولیه روز را ترجیح می‌دهند.

۲-۵- کارهای آتی

قابل ذکر است که این پژوهش مانند هر پژوهشی، محدودیت‌هایی را به همراه داشته است. محدودیت نخست این بود که تحلیل لاگ‌ها تنها بر روی ۳ ماهه سوم سال ۹۵ صورت گرفت و حجم محدودی از کل لاگ‌ها را شامل می‌شد. محدودیت دوم این بود که لاگ‌های سامانه‌ی گنج، اطلاعات کمی را ذخیره می‌کنند که خود این مسئله یافته‌های کمتری را در اختیار پژوهش قرار می‌دهد. ویژگی‌های ذخیره شده در لاگ‌ها، تنها شامل عبارت مورد جستجو، زمان جستجو، تعداد نتایج جستجو و محل جستجو است. برای پژوهش‌های آتی می‌توان به ارتباط کلیدواژه‌ها، خوشه‌بندی آن‌ها و تشکیل گراف برای تحلیل بیشتر و دقیق‌تر رفتار کاربران اشاره نمود. همچنین می‌توان ویژگی‌های بیشتری را در لاگ‌ها ذخیره نمود از جمله اینکه کلیک کاربر بر روی کدام نتایج بوده است، زمان کلیک کاربر بر روی هر نتیجه، مدت زمان حضور کاربر بر

روی هر یک از نتایج، جنسیت کاربر، مقطع تحصیلی، سن آن‌ها غیره. واضح است که با ذخیره کردن ویژگی‌های بیشتر از کاربران و رفتار آن‌ها در زمان تعامل با سامانه گنج، تحلیل‌های بیشتری می‌توان ارائه کرد.

با توجه به اینکه طرح پژوهشی فعلی بر روی سامانه گنج نسخه‌ای انجام گرفته که در زمان پیشنهاد طرح در حال استفاده بود و در طول انجام پژوهش، سامانه گنج نسخه‌ی جدید بهره برداری شد. لذا بخشی از هدف این پژوهش ارائه لیستی از فیلدهایی بود که در لاگ سامانه گنج جدید مورد استفاده قرار گیرد. لیست کامل فیلدهای پیشنهادی جهت ذخیره‌سازی لاگ‌ها در سامانه گنج جدید به شرح زیر می‌باشد:

جدول ۱۴- لیست کامل فیلدهای پیشنهادی جهت ذخیره‌سازی لاگ‌ها در سامانه گنج جدید

شماره	لاگ پیشنهادی	آماده‌سازی	توضیح
۱	زمان جست‌وجو	Timestamp (RequestLog)	
۲	عبارت جست‌وجو شده	RequestLog	
۳	جست‌وجو ساده براساس چه فیلدهایی صورت گرفته است (پدیدآورنده، موضوع و ...)	RequestLog	
۴	استفاده از جست‌وجوی پیشرفته و عملگرها (و، یا، و نه)	RequestLog	
۵	اطلاعات پروفایل کاربری به طور کامل	Log نیست.	
۶	شناسه نشست (SessionID)	UserLog	
۷	طول کوئری مورد جست‌وجو	RequestLog + comment	
۸	تعداد نتایج برگردانده شده برای هر جست‌وجو	RequestLog	
۹	تعداد نتایج مشاهده شده توسط کاربر	RequestLog	کلیک شده
۱۰	استفاده از pagination برای مشاهده نتایج		
۱۱	زمان کلیک بر روی هر نتیجه	Timestamp (RequestLog)	تقدم و تأخر
۱۲	مدت زمان حضور کاربر بر روی هر یک از نتایج	Timestamp (RequestLog)	فاصله بین دو Req، req مشاهده صفحه جزئیات تا req مستلزم خروج از صفحه.

شماره	لاگ پیشنهادی	آماده‌سازی	توضیح
۱۳	کاربر چندبرگ نخست را مشاهده کرده یا متن کامل؟ یا هر دو؟ زمان کلیک بر روی هر کدام و مدت زمان مشاهده هر کدام ابتدا کدام را کلیک کرده	RequestLog	
		Timestamp (RequestLog)	
		Timestamp (RequestLog)	فاصله بین دو Req (مانند ردیف ۱۰)
		Timestamp (RequestLog)	
۱۴	کلیک کاربر بر روی تمام آیتم‌های موجود بر روی صفحه نتایج و زمان کلیک (گزینه‌های هم‌رسانی، آراس‌اس، آگاه‌ساز، تب چکیده، تب کلیدواژه، گزارش اشکال	RequestLog	
۱۵	کلیک کاربر بر روی تمام آیتم‌های موجود بر روی صفحه جزئیات و زمان کلیک (گزینه‌های هم‌رسانی، آراس‌اس، گزارش اشکال، آمار، تب چکیده، تب کلیدواژه، منابع فارسی، منابع غیرفارسی، فهرست نوشته‌ها)	RequestLog	
۱۶	دنباله‌ی کلیک‌ها	RequestLog (ClickStream)	
۱۷	استفاده از مرتب‌سازی در صفحه نتایج	RequestLog	
۱۸	استفاده از فیلترها (کنار صفحه) در صفحه نتایج	RequestLog	
۱۹	استفاده از فیلترها (فراداده کلیک‌شدنی) در صفحه نتایج	RequestLog	
۲۰	استفاده از فیلترها (فراداده کلیک‌شدنی) در صفحه جزئیات	RequestLog	
۲۱	استفاده از پروفایل موجودیت‌ها (ارتباطات پژوهشی) در صفحه جزئیات	RequestLog	
۲۲	استفاده از امکان تغییر شمار نتایج نمایش داده شده در صفحه نتایج	RequestLog	
۲۳	استفاده از میز کار کاربری	RequestLog	

فهرست منابع

۱. جریری، نجلا؛ اسدی، مریم؛ نوشین فرد، فاطمه. ۱۳۹۲. تحلیل رفتاری جستجوی اطلاعات پژوهشگران حوزه‌های مختلف علوم از وب براساس سبک‌های شناختی کلامی و تصویری. پژوهشنامه‌ی پردازش و مدیریت اطلاعات
۲. جمالی مهموئی، حمیدرضا. ۱۳۸۳. یادداشتی بر روش تحلیل گزارش‌های وب (وبلاگ). اطلاع‌شناسی. ۱ (۳): ۱۷۵-۱۸۶
۳. خسروی، مریم؛ مهموئی، حمیدرضا. ۱۳۹۲. تحلیل لاگ پایگاه اطلاعات و مدارک علمی ایران (ایرانداک) و رفتار جستجوی کاربران آن. پژوهشنامه‌ی پردازش و مدیریت اطلاعات. پژوهشگاه علوم و فناوری اطلاعات ایران.
۴. رجبعلی بگلو، رضا. ۱۳۸۶. آشنایی با شیوه‌های تحلیل گزارش: فرایند تعامل در نظام‌های الکترونیکی. فصلنامه کتابداری و اطلاع‌رسانی. ۱۰ (۳۹).
۵. زوارقی، رسول. ۱۳۸۴. تحلیل گزارش‌های وب، روشی نوین برای ارزیابی عملکرد وب سایت‌ها: مطالعه‌ی موردی: وب ایران سایت مرکز اطلاعات و مدارک علمی. اطلاع‌شناسی ۲ (۳-۴): ۸۸-۱۱۶.
۶. ستوده، هاجر. ۱۳۸۲. "روش تحلیل گزارش‌های وب (وب لاگ). اطلاع‌شناسی. ۱ (۱): ۶۹-۸۴.

7. Alspaugh, S., Di Chen, B., Lin, J., Ganapathi, A., Hearst, M. A., & Katz, R. H. (2014, November). Analyzing Log Analysis: An Empirical Study of User Log Mining. In *LISA* (pp. 53-68).
8. Ahiauzu, B. E., & Ani, O. E. (2015). A Survey of Information Seeking-Behavior of Academic Staff in a Nigerian University in Digital Age. *Science*, 3(4), 89-94.
9. Baeza-Yates, R., Hurtado, C., Mendoza, M., & Dupret, G. (2005). Modeling user search behavior. In *Web Congress, 2005. LA-WEB 2005. Third Latin American* (pp. 10-pp). IEEE.
10. Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., & Frieder, O. (2004, July). Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 321-328). ACM.
11. Belkin, N. J. (1978). Information concept for Information Science. *Journal of Documentation* 34 (1), 55-58.
12. Bendersky, M., & Croft, W. B. (2009, February). Analysis of long queries in a large scale search log. In *Proceedings of the 2009 workshop on Web Search Click Data* (pp. 8-14). ACM.
13. Broder, A. (2002, September). A taxonomy of web search. In *ACM Sigir forum* (Vol. 36, No. 2, pp. 3-10). ACM.
14. Cetindil, I., Esmaelnezhad, J., Li, C., & Newman, D. (2012). Analysis of Instant Search Query Logs. In *WebDB* (pp. 7-12).
15. C.S. Turnor. (1998) *Organizing Information principles & practice*. Clive Bingley: London, 217-255.
16. Dervin and M. Milan. (1986) *Information needs and uses in: Annual Review of Information Science and Technology*. Knowledge Industry Publication: New York, 3-33.
17. Downey, D., Dumais, S., Liebling, D., & Horvitz, E. (2008, October). Understanding the relationship between searchers' queries and information goals. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 449-458). ACM.
18. Dupret, G. E., & Piwowarski, B. (2008, July). A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 331-338). ACM.

19. Ellis, D., & Oldman, H. (2005). The English literature researcher in the age of the Internet. *Journal of Information Science*, 31(1), 29-36.
20. Facca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from weblogs: a survey. *Data & Knowledge Engineering*, 53(3), 225-241.
21. Gan, Q., Attenberg, J., Markowetz, A., & Suel, T. (2008, April). Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web* (pp. 49-56). ACM.
22. Gleeson, A. C. (2001). Information-seeking behavior of scientists and their adaptation to electronic journals. *University of North Carolina dissertation*.
23. Ingwersen, P. (2008). Information and Information Science in context. In J. Olaisen, E. Munch-Petersen, P. Wilson (eds.). *Information Science. From the development of the discipline to social interaction*. Oslo: Scandinavian University Press pp 69-111.
24. Jansen, B. J., Taksa, I., & Spink, A. (2009). Handbook of Research on Web Log Analysis, chapter Research and Methodological Foundations of Transaction Log Analysis. *IGI Global*, 226-253.
25. Jansen, B. J. (2009). Understanding user-web interactions via web analytics. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1), 1-102.
26. Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information processing & management*, 36(2), 207-227.
27. Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information processing & management*, 42(1), 248-263.
28. Jiandong, W., Mingzi, L., & Jimin, W. A Study on Chinese University Users' Academic Search Behavior Based on Log Mining.
29. Kumar, A., Salmani, N., & Baweja, S. (2014). Information Seeking Behaviour by the Research Scholars & Faculty Members: A Survey Study of Kurukshetra University Kurukshetra in the Disciplines of Life Science. *IOSR Journal of Humanities and Social Science (IOSR-JHSS)*, 19(6), 119-138.
30. Kurth, M. (1993). The limits and limitations of transaction log analysis. *Library Hi Tech*, 11(2), 98-104.
31. Liawa, S. S., & Huang, H. M. (2003). An investigation of user attitudes toward search engines as an information retrieval tool. *Computers in human behavior*, 19(6), 751-765.
32. Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science*, 40(1), 54.
33. Meister, D., & Sullivan, D. J. (1967). Evaluation of User Reactions to a Prototype On-line Information Retrieval System: Report prepared under Contract No. NASw-1369 by Bunker-Ramo Corporation. Report Number NASA CR-918. Oak Brook, IL: Bunker-Ramo Corporation. Document Number N67-40083).
34. Morzy, T., Wojciechowski, M., & Zakrzewicz, M. (2000). Web users clustering. In *Proc. of the 15th International Symposium on Computer and Information Sciences* (pp. 374-382).
35. Murray, G. C., & Teevan, J. (2007, December). Query log analysis: social and technological challenges. In *ACM SIGIR Forum* (Vol. 41, No. 2, pp. 112-120). ACM.
36. Nicholas, D., Huntington, P., & Jamali, H. R. (2008). User diversity: as demonstrated by deep log analysis. *The Electronic Library*, 26(1), 21-38.

37. Nielsen Media. (1997). Search engines most popular method of surfing the web [Website]. Commerce Net/Nielsen Media. Retrieved 30.8.2000 from the World Wide Web: <http://www.commerce.net/news/press/0416.html>.
38. Oduntan, O. (2014). Information seeking behavior on an academic library search engine.
39. Ozmutlu, S., Ozmutlu, H. C., & Spink, A. (2003). Are people asking questions of general Web search engines?. *Online Information Review*, 27(6), 396-406.
40. Park, M., & Lee, T. S. (2013). Understanding science and technology information users through transaction log analysis. *Library Hi Tech*, 31(1), 123-140.
41. Pareek, A. K., & Rana, M. S. (2013). Study of information seeking behavior and library use pattern of researchers in the Banasthali University.
42. Peters, T. A. (1993). The history and development of transaction log analysis. *Library hi tech*, 11(2), 41-66.
43. Rose, D. E., & Levinson, D. (2004, May). Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web* (pp. 13-19). ACM.
44. Shivalingaiah, D. (2010). Information seeking patterns of researchers in the university libraries in Karnataka state. *SRELS Journal of Information Management*, 47(1), 83-101.
45. Silvestri, F. (2009). Mining query logs: Turning search usage data into knowledge. *Foundations and Trends® in Information Retrieval*, 4(1-2), 1-174.
46. Spink, A., & Jansen, B. J. (2006). Searching multimedia federated content web collections. *Online Information Review*, 30(5), 485-495.
47. Tyler, S. K., & Teevan, J. (2010, February). Large scale query log analysis of re-finding. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 191-200). ACM.
48. Wang, P., Berry, M. W., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. *Journal of the Association for Information Science and Technology*, 54(8), 743-758.
49. Wen, J. R., Nie, J. Y., & Zhang, H. J. (2002). Query clustering using user logs. *ACM Transactions on Information Systems*, 20(1), 59-81.
50. White, R. W., & Morris, D. (2007, July). Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 255-262). ACM.
51. Wiberley, S. E., & Jones, W. G. (1989). Patterns of information seeking in the humanities. *College & Research Libraries*, 50(6), 638-645.
52. Wilson, T. D. (2000). Human information behavior. *Informing science*, 3(2), 49-56.
53. Xie, K., Yu, H., & Cen, R. (2012). Using log mining to analyze user behavior on search engine. *Frontiers of Electrical and Electronic Engineering*, 7(2), 254-260.
54. Xi, W., Chowdhury, A., Sidhu, K., & Pass, G. (2001). Query Log Analysis. *Laporan teknis, American Online, Inc.*
55. Yi, K., Beheshti, J., Cole, C., Leide, J. E., & Large, A. (2006). User search behavior of domain-specific information retrieval systems: An analysis of the query logs from PsycINFO and ABC-Clio's Historical Abstracts/America: History and Life. *Journal of the Association for Information Science and Technology*, 57(9), 1208-1220.
56. Yu, L., & Apps, A. (2000). Studying e-journal user behavior using log files: the experience of superjournal. *Library & Information Science Research*, 22(3), 311-338.
57. Ke, H. R., Kwakkelaar, R., Tai, Y. M., & Chen, L. C. (2002). Exploring behavior of E-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. *Library & Information Science Research*, 24(3), 265-291.

58. Park, M., Park, M., Lee, T. S., & Lee, T. S. (2016). A longitudinal study of information needs and search behaviors in science and technology: A query analysis. *The Electronic Library*, 34(1), 83-98.
59. Rabiei, M., Hosseini-Motlagh, S. M., & Haeri, A. (2017). Using text mining techniques for identifying research gaps and priorities: a case study of the environmental science in Iran. *Scientometrics*, 110(2), 815-842.
60. Sato, S., Nishiura, M., Nagai, Y., Itsumura, H., サトウショウ, ニシウラミナコ, ... & 永井裕子. (2015). Usage Log Analysis of Articles in Five Japanese Institutional Repositories: the Relationships between Users, Access Paths, and Accessed Articles. *同志社図書館情報学*= *Doshisha Journal of Library and Infomation Science*, (25), 20-37.

Abstract.

One of the most important goals of each information system is to provide the best services according to users' needs. This increases user satisfaction and they will be interested in using information systems. Nowadays, search engines are one of the most important tools for data retrieval and they provide a variety of information for many users. Students, university professors, and researchers search daily in search engines and meet their information needs. Ganj database of the Iranian Science and Technology Research Institute has 870,000 scientific records which allow researchers to search for scientific publications, articles, conferences' papers, research projects' reports, and government reports. It is clear that users are influenced by different factors when they are searching. They use various methods for obtaining information. In addition, they have different behavior search. The main purpose of the present research is to analyze the behavior of users' search information in the Ganj search engine. Users' behavior analysis can help to better understand users' needs.

The results of the analysis indicate that the using Ganj in the Iranian Science and Technology Research Institute is very small in comparison with the using Ganj out of the Iranian Science and Technology Research Institute. Also, the hours of the search are different among the researchers in the field of science and technology. These results of help to assess and resolve possible deficiencies in the field of information retrieval by the search engine.

Key words: Information-Seeking behavior, behavior analysis, search engine, information system, Ganj System



Ministry of Science Research and Technology

**Iranian Research Institute
for Information Science and Technology (Irandoct)**

**Analyzing researchers' seeking information behavior on
the GANJ search engine**

Investigator:

Somayeh Fatahi

December 2017